ستيوارت راسل

ذكاء اصطناعي متوافق مع البشر

حتى لا تفرض الآلات سيطرتها على العالم



ترجمة مصطفى محمد فؤاد وأسامة إسماعيل عبد العليم

ذكاء اصطناعي متوافق مع البشر

حتى لا تفرض الآلات سيطرتها على العالم

تأليف
ستيوارت راسل

ترجمة
أسامة إسماعيل عبد العليم



ذكاء اصطناعي متوافق مع البشر

ستيوارت راسل

Human Compatible

Stuart Russell

الناشر مؤسسة هنداوي
المشهرة برقم ١٠٥٨٥٩٧٠ بتاريخ ٢٦ / ١ / ٢٠١٧

يورك هاوس، شييت ستريت، وندسور، SL4 1DD، المملكة المتحدة
تليفون: ٨٣٢٥٢٢ ١٧٥٣ (٠) ٤٤ +
البريد الإلكتروني: hindawi@hindawi.org
الموقع الإلكتروني: https://www.hindawi.org

تصميم الغلاف: يوسف غازي

الترقيم الدولي: ٩ ٢٩٢٤ ٥٢٧٣ ١ ٩٧٨

صدر الكتاب الأصلي باللغة الإنجليزية عام ٢٠١٩.
صدرت هذه الترجمة عن مؤسسة هنداوي عام ٢٠٢٢.

المحتويات



إهداء إلى لوي وجوردون ولوسي وجورج وإيزاك.

شكر وتقدير

لقد ساهم الكثير من الأشخاص في خروج هذا الكتاب للنور. من بين هؤلاء محرِّراي المُتميزان في دار نشر فايكنج (بول سلوفاك) ودار نشر بنجوين (لورا ستيكني)؛ ووكيلي جون بروكمان، الذي حثَّني على تأليف شيء في هذا الموضوع؛ وجيل ليوفي وروب ريد، اللذان قدَّما لي الكثير من التعليقات المفيدة؛ والمراجعون الآخرون للبروفات الأولى للكتاب، وبخاصة زياد مرار ونك هاي وتوبي أورد وديفيد ديوفنود وماكس تيجمارك وجريس كاسي. وقد ساعدتْني كارولين جانمير بشدة في فحص الاقتراحات الهائلة الخاصة بالتحسينات التي قدمها مُراجعو البروفات الأولى، في حين تولى مارتن فوكوي مهمة جمع التراخيص الخاصة بعرض الصور.

الأفكار التقنية الأساسية المعروضة في الكتاب جرى تطويرها بالتعاون مع أعضاء مركز الذكاء الاصطناعي المُتوافق مع البشر بجامعة كاليفورنيا في بيركلي، وبخاصة توم جريفيث وأنكا دراجان وأندرو كريتش وديلان هادفيلد-مينيل وروبن شاه وسميثا ميلي. يقود هذا المركز على نحو رائع المدير التنفيذي مارك نتزبرج والمدير المُساعد روزي كامبل، وتُموِّله على نحوٍ سخيٍّ مؤسَّسة أوبن فيلانتربي.

وقد ساعدت رامونا ألفاريز وكارين فيردو في جعل الأمور تسير بسلاسة طوال فترة تأليف الكتاب، ومنحتْني زوجتي الرائعة، لوي، وأبنائي؛ جوردون ولوسي وجورج وإيزاك، قدرًا هائلًا وضروريًّا من الحب والصبر والتشجيع لإكماله، لكن ليس دائمًا بهذا الترتيب الموضَّح.

مقدمة

(١) لماذا كُتب هذا الكتاب؟ ولم كُتب الآن؟

يبحثُ هذا الكتابُ مُحاولاتنا لفهم ماهيَّة الذكاء ومساعينا لُمضاهاته في الماضي والحاضر والمستقبل. وهذا موضوع جوهري؛ لا لأنَّ الذكاء الاصطناعي قد تطوَّر بسرعة ليصير مظهرًا سائدًا من مظاهر حاضرنا، بل لأنَّه سيكون وبلا شكٍّ التقنية المُهيمنة في مُستقبلنا. إننا نرى القُوى العُظمى في العالم تستفيق أخيرًا لتُدرك هذه الحقيقة، كما نُلاحظُ أنَّ أكبر شركات العالم قد انتبهت لها منذ عدة سنوات. ونحن إذ لا نستطيع أن نتنبَّأ بدقةٍ بكيفية تطوُّر هذه التِّقنية ولا وفق أيِّ جدولٍ زمني، فإنني أرى أنَّ لزامًا علينا أن نُخطِّط لاحتمالية أن تتخطَّى الآلات مقدرة البشر العقلية على اتِّخاذ القرارات في العالم الواقعي. فما الذي سنفعله حينها؟

كلُّ ما في جعبة الحضارة الإنسانية وما وصلت إليه هو نتاجُ ذكائنا البشري، أما أن نضع أيدينا على مصدر ذكاءٍ أعظم بكثيرٍ مما لدينا، فسيكون هذا حدثًا فارقًا في تاريخ البشرية. غاية هذا الكتاب أن يُفسِّر لماذا قد يكون ذلك الحدث هو آخر أحداث التاريخ البشري، وكيف نحرص على ألَّا يكون كذلك.

(٢) عرض مُوجز لُمحتوى الكتاب

ينقسم هذا الكتاب إلى ثلاثة أجزاء. يستطلع الجزء الأول، في فصولٍ ثلاثة، مفهوم الذكاء في البشر وفي الآلات. لا تتطلَّب المادة المعروضة منك أي سابق معرفةٍ بالمجال التِّقني، لكن إن كنت ذا اهتمامٍ بالمجال، فالكتاب مُذيل بأربعة ملاحق يشرح كلُّ واحدٍ منها بعض المفاهيم

الأساسية التي ترتكزُ عليها نظم الذكاء الاصطناعي الحالية. أما الجزء الثَّاني، فيُناقشُ في ثلاثة فصولٍ بعض المشاكل التي نجمت عن غرس الذكاء في الآلات. وأُركِّز فيه تحديدًا على «مُعضلة التَّحكُّم»؛ وهي كيف نُبقي على تحكُّمٍ مُطلقٍ بالآلات التي أصبحت أقوى منَّا. أما الجزء الثالث والذي يمتدُّ على مدى أربعة فصول، فيقترحُ طريقةً جديدة للنظر إلى الذكاء الاصطناعي ولضمان أن تظلَّ الآلات في خدمة البشر إلى الأبد. جدير بالذِّكر أنَّ هذا الكتاب يستهدف عامَّة القُرَّاء، لكني آمُلُ أن يكون ذا نفعٍ في حمل المُتخصِّصين في مجال الذكاء الاصطناعي على الاقتناع بإعادة التَّفكير في فرضيَّاتهم الأساسيَّة.

الفصل الأول

ماذا لو نجحنا؟

مُنذ فترة طويلة، كان والداي يعيشان في مدينة برمنجهام بإنجلترا في بيتٍ قُرب الجامعة. في يومٍ من الأيام، قرَّرا الرَّحيل عن المدينة وباعا المنزل إلى ديفيد لودج؛ أستاذ الأدب الإنجليزي. حينها، كان ديفيد في أوج مجده وشُهرته كروائي. ومع أنِّي لم أقابله قط، لكني عقدت العزم على قراءة بعضٍ من كُتُبه؛ على سبيل المثال، رواية «تبادل الأماكن» ورواية «عالم صغير». ومن بين الشخصيات الرئيسية في هاتين الروايتين، هناك بعض الأكاديميِّين الخياليِّين الذين ينتقلون من نُسخةٍ خياليةٍ لمدينة برمنجهام إلى نُسخةٍ خياليةٍ لمدينة بيركلي بولاية كاليفورنيا. وعندما كنتُ أنا أكاديميًّا حقيقيًّا من مدينة برمنجهام الحقيقية وقد انتقلتُ لتوِّي إلى مدينة بيركلي الحقيقية، شعرتُ حينها أنَّ هذه مُصادفة لا يجدُر بي أن أدعها تمرُّ مرور الكرام دون تمعُّنٍ وانتباه.

لقد أعجبني أحد المشاهد في رواية «عالم صغير»؛ حيث كان بطل الرواية، الذي كان باحثًا أدبيًّا طموحًا، يحضُر مؤتمرًا عالميًّا مُهمًّا ثمَّ يسأل لجنة المُناقشة التي تضُمُّ عددًا من الشخصيات القياديَّة: «ما خطوتكم التالية إذا وافقكم الجميع الرأي؟» أثار ذلك السؤال فزعًا وذُعرًا لأن المُناقشين كانوا مُنهمكين في الصراع الفكري بدلًا من تحرِّي الحقائق ومحاولة الوصول إلى فهمٍ صحيح. وحينها، طرأ سؤال مُشابه في ذهني أريد أن تُجيب عنه الشخصيات القيادية في مجال الذكاء الاصطناعي: «لنفترض أنكم نجحتم، ماذا بعد؟» إن الهدف الأسمى لهذا المجال كان ولا يزال خلق ذكاءٍ اصطناعي يُماثلُ الذكاء البشري أو يفُوقُه. ولكنَّنا لم نفكر، اللَّهُم إلا من بعض المحاولات المُتواضعة، فيما سيئول إليه الحال لو نجحْنا في مسعانا ذاك.

بعد سنواتٍ قليلةٍ، بدأتُ أنا وبيتر نورفج تأليف كتابٍ جديد عن الذكاء الاصطناعي، ونُشِرت أول طبعةٍ منه عام ١٩٩٥.[1] وكان عنوان آخر قسم فيه هو «ماذا لو فعلناها ونجحنا؟» وكان ذاك القسم يُشير إلى العواقب الحسنة والسَّيئة واحتمالاتهما، لكنَّه لم يصل إلى استنتاجاتٍ مُحكمة. وحين صدرت الطَّبعة الثالثة من الكتاب عام ٢٠١٠، كان الكثير من النَّاس قد بدءوا يأخذون بعين الاعتبار احتمالية أنَّ الذكاء الاصطناعي الخارق قد لا يكون أمرًا جيدًا؛ ولكن كان مُعظم هؤلاء غير مُتخصِّصين وليسوا من عُمُوم الباحثين في مجال الذكاء الاصطناعي. وفي عام ٢٠١٣، وصلتُ إلى قناعةٍ أنَّ تلك المسألة لا تخُصُّ مُجتمع الباحثين في المجال فقط، بل رُبَّما كانت أعظم تساؤلٍ يُواجه البشرية جمعاء.

في شهر نوفمبر عام ٢٠١٣، ألقيتُ محاضرةً في معرض صور داليتش؛ وهو معرض فنِّي عريق جنوب لندن. كان مُعظم الحضور من المُتقاعدين غير المُتخصِّصين، ولكنَّهم كانوا ذوي اهتمامٍ عامٍّ بالقضايا الفكرية. لذلك كان عليَّ أن أُلقي محاضرةً مُبسَّطةً تمامًا، وقد بدا هذا المقامُ ملائمًا لأطرح فيها أفكاري على الملأ للمرة الأولى. وهكذا، بعد أن شرحتُ ما هو الذكاء الاصطناعي، رشَّحتُ خمسة أحداثٍ ليكون أحدها هو «أعظم حدثٍ في مستقبل البشرية»:

(١) أن نهلك جميعًا (سواء بارتطامٍ نيزكي أم كارثةٍ مناخيَّةٍ أم تفشٍّ لوباءٍ خطير وهلُمَّ جرًّا).
(٢) أن نعيش مُخلَّدين للأبد (باكتشاف إكسير الحياة والحدِّ من الشيخوخة).
(٣) أن نخترع السَّفر بسرعةٍ تفوقُ سرعة الضَّوء ونغزو الكون.
(٤) أن تغزُونا كائنات فضائية من حضارةٍ أكثر تطورًا من حضارتنا.
(٥) أن نخترع ذكاءً اصطناعيًّا خارقًا.

وتوقَّعت حينها أن يكون الحدث الخامس؛ الذكاء الاصطناعي الخارق، هو الفائز. فهو سيساعدنا على تجنُّب الكوارث المادِّية وتحقيق الخُلُود واختراع السَّفر بسرعةٍ تفوقُ سرعة الضَّوء، إن كانت هذه الأشياء مُمكنة الحدوث أصلًا. كما سينقل حضارتنا البشرية نقلةً كبيرةً، بل قد يخلق حضارةً جديدة تمامًا. فاليوم الذي نخترعُ فيه ذكاءً اصطناعيًّا خارقًا سيكون مُماثلًا، من نواحٍ كثيرة، لليوم الذي تصل فيه كائنات فضائية من حضارةٍ أكثر تطورًا من حضارتنا إلى كوكبنا، لكنَّه في الغالب هو الأقرب للحدوث. ورُبَّما كان أهم ما في الأمر أنَّ الذكاء الاصطناعي هو شيء نملكُ زمامه إلى حدٍّ ما، على عكس الكائنات الفضائية.

بعدها، طلبت من الجمهور أن يتخيَّلوا ماذا سيحدث إذا تلقَّينا إنذارًا من كائناتٍ فضائيةٍ من حضارةٍ مُتفوِّقةٍ يُخبروننا فيه أنَّهم سيقدُمُون إلى كوكب الأرض في غضون الثلاثين إلى الخمسين سنةً المُقبلة؟ دعوني أُخبركم أنَّ القاعة امتلأت بالهرج والمرج. ولكن يكفي أنْ أقول إن ردَّ فعلهم على توقُّعات اختراع ذكاء اصطناعي خارق كان أقل من المتوقع. (في محاضرةٍ لاحقةٍ، وضَّحتُ ذلك في صورة مُراسلة بريد إلكتروني ترونها في الشَّكل ١-١.) وأخيرًا، أوضحتُ لهم مدى أهمية الذكاء الاصطناعي الخارق وخُطُورته في الوقت ذاته، فقلتُ: «نجاحنا في هذا الأمر سيكون أعظم حدثٍ في تاريخ البشرية … ورُبَّما آخر أحداثها على الإطلاق.»

مِن: كائنات فضائية من حضارةٍ مُتفَوِّقة <sac12@sirius.canismajor.u>
إلى: البَشَرية <humanity@UN.org>
الموضوع: رسالة تواصُل
خذوا حذركم! سنصل إلى كوكبكم في غضون سنين؛ من ٣٠ إلى ٥٠ سنة.

مِن: البَشَرية <humanity@UN.org>
إلى: كائنات فضائية من حضارةٍ مُتفَوِّقة <sac12@sirius.canismajor.u>
الموضوع: معذرةً! نحن في عُطلَة. ردًّا على: رسالة تواصُل
البشرية حاليًّا في إجازة. سنَرُدُّ على رسالتكم عندما نعود.☺

شكل ١-١: ربما لا تكون هذه هي المُراسلة البريدية التي ستنتُج عن أول تواصلٍ مع حضارةٍ فضائية متفوِّقة.

مرَّت عدة أشهرٍ، وبالتَّحديد في شهر أبريل عام ٢٠١٤، كنتُ أحضُر مؤتمرًا في أيسلندا عندما تلقَّيت اتصالًا من «الإذاعة الوطنية العامة» يسألونني فيه إذا كنتُ أودُّ أن أجري حوارًا نقاشيًّا حول فيلم «التَّسامي» («ترانسندنس»)؛ الذي كان قد بدأ عرضُهُ حديثًا في الولايات المتحدة. كنتُ قد قرأتُ عددًا من مُلخَّصات حبكة الفيلم وبعض مُراجعاتٍ له، لكنِّي لم أشاهده لأني كنت أعيش في باريس وقتها، ولم يكن ليُعرض هناك إلا في شهر يونيو. ثمَّ اضطُررتُ أن أعرِّج على مدينة بوسطن في طريقي من أيسلندا إلى بيتي لأشارك في اجتماعٍ لوزارة الدِّفاع. وهكذا وفور أن وصلتُ إلى مطار لوجان الدولي بمدينة بوسطن، ركبت سيارة أجرةٍ إلى أقرب دار سينما تعرض الفيلم، ثمَّ جلستُ في الصَّف الثاني وشاهدت

المُمثِّل جوني ديب، في دور أستاذ ذكاءٍ اصطناعي ببيركلي، وهو يُواجه محاولة اغتيالٍ من ناشطين مُعادين للذكاء الاصطناعي، وهم، كما جال في خاطرك، جماعة تخشى عواقب الذكاء الاصطناعي الخارق. حينها، انكمشتُ في مقعدي لا إراديًّا. (أهذه مصادفة أخرى يجب أن أقف عندها لأُراجع نفسي؟) وقبل موت الشخصية التي يُجسِّدها جوني ديب، حُمل عقله إلى كمبيوترٍ كمِّي فائق السُّرعة، ثمَّ ما لبث أن صار ذا قدراتٍ تتخطَّى حدود القدرات البشرية وبدأ يُهدِّد بالسيطرة على العالم.

وفي التاسع عشر من شهر أبريل عام ٢٠١٤، نشرتُ مُراجعةً للفيلم على موقع «هافينجيتون بوست» بالمُشاركة مع الفيزيائيِّين ماكس تيجمارك، وفرانك ويلتشك، وستيفين هوكينج. تضمَّنت المُراجعة الجملة التي قلتُها في محاضرة معرض داليتش عن أعظم حدثٍ في تاريخ البشرية. ومنذ ذلك الحين، تبنَّيت علنًا وجهة النَّظر القائلة بأنَّ مجال بحثي قد يُشكِّلُ تهديدًا مُحتملًا لأبناء جنسي البشري.

(١) كيف وصلنا إلى هنا؟

بدأ البحث في مجال الذكاء الاصطناعي منذ فترةٍ طويلة، لكنَّ بدايته «الرسمية» تُؤرَّخ بعام ١٩٥٦ عندما أقنع جون ماكارثي ومارفن مينيسكي؛ وهما عالما رياضيَّاتٍ شابَّان، كُلًّا من كلود شانون الذي كان وقتها مشهورًا بصفته مُخترع نظرية المعلومات، وناثانيل رتشيستر؛ مُصمِّم أول كمبيوتر يُباع في الأسواق من شركة آي بي إم، أن ينضمَّا إليهما لتنظيم برنامج صيفي في جامعة دارتموث. وكان الهدف منه كما يلي:

> يقوم البرنامج على افتراض أنَّ كل جوانب التَّعلُّم أو أي سمةٍ من سمات الذكاء يُمكن، نظريًّا، أن تُوصَّف توصيفًا دقيقًا بحيث يُمكن جعل الآلات قادرةً على محاكاتها. ستُجرى محاولة لاكتشاف كيفية جعل الآلات تتحدَّثُ اللغة؛ وتصُوغ الأفكار المُجرَّدة والمفاهيم؛ وتعمل على حلِّ ذاك الضَّرب من المشاكل المُستعصية والمقصُور البحث فيها على البشر؛ وتُطوِّر من نفسها. نظُنُّ أنَّ تقدُّمًا ملحوظًا يُمكن أن يُحرز في واحدةٍ أو أكثر من هذه المسائل إذا ما اشتغل بها فريق من العلماء مُنتقًى بعنايةٍ خلال صيفٍ واحد.

لا حاجة بنا للإشارة إلى أنَّ تلك التجربة قد استغرقت وقتًا أطول بكثيرٍ من فصل صيفٍ واحد؛ فنحن ما نزال إلى الآن نعمل على حلولٍ لتلك المسائل.

في خلال العقد الأول أو نحو ذلك بعد برنامج دارتموث، ازدهر الذكاء الاصطناعي وشهد العديد من النَّجاحات الهامَّة؛ بما في ذلك خوارزمية آلان روبنسون للتفكير المنطقي العام[2] وبرنامج لُعبة الدَّامة الذي صمَّمه آرثر صامويل، والذي طور من نفسه حتى تغلَّب على صانعه.[3] أما أول فقاعةٍ للذكاء الاصطناعي، فقد انفجرت في أواخر السِّتينيَّات من القرن العشرين، عندما فشلت الجهود المُبكِّرة في مجالَي تعلُّم الآلة والتَّرجمة الآلية في الارتقاء إلى مستوى التوقعات. وخلُص تقرير أعدَّته الحكومة البريطانية عام ١٩٧٣ إلى أنَّنا «لا نستطيع أن نُشير إلى أي فرعٍ من فروع هذا المجال ونقول إن الاكتشافات التي أحرزت فيه حتى الآن قد حقَّقت الأثر الهائل الذي كان متوقَّعًا منها.»[4] أو بعبارةٍ أخرى، لم تكن الآلات ذكيةً بما يكفي.

عندما كنت في سنِّ الحادية عشرة، لحُسن حظِّي، لم أكن أعرف شيئًا عن هذا التَّقرير. وبعد سنتين، أُهديت إليَّ آلة حاسبة قابلة للبرمجة من طراز «سينكلير كامبريدج»، وحينها أردت فقط أن أجعلها ذكية. ولكن تلك الآلة الحاسبة التي ما كانت ذاكرتها لتحتمل أكثر من ٣٦ خطوة حسابية، لم تكن كبيرةً كفايةً بحيث تمتلك ذكاءً اصطناعيًّا يُضاهي الذَّكاء البشري. بعدها، وأنا غير مُهبَّط العزيمة، تمكنت من الوصول إلى الكمبيوتر الفائق «سي دي سي ٦٦٠٠»[5] ذي الحجم الضَّخم، في كلية إمبريال كوليدج بلندن، وأنشأتُ عليه برنامج لُعبة شطرنج، والذي كان مُخزَّنًا على مجموعةٍ من البطاقات المثقوبة التي يبلغ ارتفاعها قدمَين. لم تكن النَّتيجة مُرضيةً جدًّا، ولكن لم يُهمَّني ذلك؛ فقد كنت أعرف حينها ما الذي أريد فعله.

أصبحتُ أستاذًا في جامعة بيركلي بحُلُول مُنتصف الثمانينيات في القرن العشرين، وكان الذكاء الاصطناعي حينها يشهد صحوةً وانتعاشًا بفضل الإمكانات التجارية لما كان يُدعى بالنظم الخبيرة. وهنا كان ثاني انفجارات فُقَّاعات الذكاء الاصطناعي؛ حين فشلت هذه النظم وأثبتت عدم أهليتها للعديد من المهام التي وُكلت إليها. مرة أخرى، لم تكن الآلات ذكيةً بما يكفي. تبع ذلك شتاء طويل لم تسطع فيه شمس على الذكاء الاصطناعي، وانكمش عدد الطلَّاب في دورة الذكاء الاصطناعي التي أُدرِّسُها من حوالي ما يربو على تسعمائة طالبٍ إلى خمسةٍ وعشرين طالبًا فقط في عام ١٩٩٠.

وهنا تعلَّم مجتمع الذكاء الاصطناعي الدَّرس، وفطن إلى أنَّ الآلات يجب أن تكون أذكى، ولكن كان علينا أن نجتهد ونكدَّ في الدِّراسة لنجعل هذا الأمر مُمكنًا. فتعمَّق المجال في علم الرياضيات، ووطَّد أواصره مع فروع المعرفة العريقة كعلم الاحتمالات والإحصاء

ونظرية التَّحكم. وغُرست بُذُور النَّجاحات التي نراها اليوم خلال أيام ذلك الشِّتاء الذي خيَّم على مجال الذكاء الاصطناعي، بما في ذلك الدراسات الأوَّلية على نظم التفكير الاحتمالي الواسع النِّطاق، التي سُمِّيت فيما بعد بـ «التَّعلُّم المُتعمِّق».

وبداية من عام ٢٠١١ تقريبًا، بدأت تقنيات التَّعلُّم المُتعمِّق في إحراز نجاحاتٍ هائلةٍ في ثلاثٍ من أهمِّ المسائل غير المحسُومة في المجال؛ تمييز الكلام، وتمييز العناصر المرئية، والتَّرجمة الآلية. وإلى حدٍّ ما، الآلات في يومنا هذا تُضاهي القدرات البشرية في تلك الأمور، بل وتتفوَّق عليها أحيانًا. ففي عامي ٢٠١٦ و٢٠١٧، هزم برنامج «ألفا جو»، الذي طورته شركة ديب مايند، بطل العالم السَّابق في لعبة جو؛ لي سيدول، وبطل العالم الحالي؛ كي جيه، وهو حدث تنبَّأ بعض الخبراء أننا لن نراه يحدُث أبدًا، وإن حصل فلن يكون قبل عام ٢٠٩٧.[6]

وها نحن الآن نشهد الذكاء الاصطناعي وهو يظهر في أخبار الصَّفحات الأولى من التَّغطيات الإعلامية كل يومٍ تقريبًا. فقد أُسِّست الآلاف من الشركات النَّاشئة التي يدعمها سيل عارم من التمويلات الاستثمارية. ودرس الملايين من الطلاب دوراتٍ في الذكاء الاصطناعي وتعلُّم الآلة عبر الإنترنت، وصار الخبراء في المجال يتقاضون رواتب بملايين الدولارات. ونذكُر هنا أنَّ الاستثمارات التي تضُخُّها الصَّناديق الاستثمارية والحكومات الوطنية والشركات الكبرى تصل إلى عشرات المليارات من الدولارات سنويًّا؛ أي إن الأموال التي استُثمرت في السنوات الخمس الماضية هي أكثر مما أُنفق على المجال منذ أن بدأ. ومن المُتوقَّع أن تترك التقنيات التي ما تزال في حيِّز التطوير؛ كالسَّيارة الذاتية القيادة والمُساعد الشَّخصي الذكي، أثرًا جوهريًّا في عالمنا خلال العقد القادم. أما المنافع الاقتصادية والاجتماعية المُحتملة التي قد نجنيها من وراء الذكاء الاصطناعي فهي كثيرة ومُتعدِّدة، مما يُعطي زخمًا عظيمًا لمؤسَّسات أبحاث الذكاء الاصطناعي.

(٢) ما الخطوة التالية؟

أيعني هذا التَّقدم السريع والمُتلاحق أنَّنا على وشْك أن تسبقنا الآلات وتتخطانا؟ الإجابة هي لا؛ فهناك العديد من الطَّفرات التقنية التي يجب أن تحدُث أولًا قبل أن نشهد ميلاد آلات ذات ذكاءٍ خارق يفوق الذكاء البشري.

من المعروف أنَّ التَّنبؤ بالطَّفرات العلمية أمر غاية في الصعوبة. ولنُدرك مدى صعوبة الأمر، فلنُلق نظرةً على تاريخ أحد المجالات الأخرى التي بإمكانها أن تُبيد الحضارة الإنسانية وتقضي عليها؛ ألا وهو الفيزياء النَّووية.

في السنوات الأولى من القرن العشرين، لعلَّ أكثر الفيزيائيِّين النَّوويين شهرةً وبُرُوزًا كان العالم إرنست رذرفورد؛ مُكتشف البروتونات والرجل الذي «شطر الذرة» (انظر الشكل ١-٢«أ»). وكغيره من أرباب المجال، كان يعرف أنَّ نواة الذرة تختزن كمًّا هائلًا من الطاقة، لكنَّ الاعتقاد السَّائد حينها كان هو أنَّ الوصول إلى هذه الطاقة والانتفاع بها هو ضرب من ضُرُوب المُستحيل.

في الحادي عشر من شهر سبتمبر عام ١٩٣٣، عقدت الجمعية البريطانية لتقدُّم العلوم اجتماعها السَّنوي بمدينة ليستر. وألقى اللورد رذرفورد خطاب الجلسة المسائية. وكما فعل مرَّاتٍ عديدة في الماضي، أوهن بخطابه عزم احتمالات إنتاج الطاقة النَّووية وقال: «أي شخصٍ ينشُد مصدرًا للطاقة من تحوُّل الذرات فهو كمن يلاحق سرابًا في فلاة.» وفي الصَّباح التالي نقلت جريدة «ذا تايمز» اللندنية خطاب رذرفورد (انظر الشكل ١-٢«ب»).

تَحوُّل العناصر

نقلًا عن مراسلينا،
ليستر، ١١ سبتمبر.
سأل اللورد رذرفورد في النهاية:
ما هي الاحتمالات بعد ٢٠ أو ٣٠ سنة؟
إنها طريقة رديئة وغير فعَّالة لإنتاج الطاقة،
وأي شخصٍ يَنشُد مصدرًا للطاقة
من تَحَوُّل الذرات فهو كَمَن ينشُدُ سَرابًا في فَلاة.



(أ) (ب) (ج)

شكل ١-٢: (أ) عالم الفيزياء النووية اللورد رذرفورد. (ب) مقتطفات من تقريرٍ صحفي أعدَّته جريدة «ذا تايمز» بتاريخ ١٢ سبتمبر، ١٩٣٣، حول الخطاب الذي ألقاه رذرفورد مساء اليوم السَّابق. (ج) عالم الفيزياء النووية ليو سيلارد.

ليو سيلارد (انظر الشكل ١-٢«ج»)، وهو فيزيائي مجري هرب من جحيم ألمانيا النَّازية، كان يُقيم في فندق إمبريال في ميدان راسل بلندن. قرأ تقرير جريدة «ذا تايمز»

وهو يتناول فطوره. ثمَّ ذهب إلى نُزهةٍ على قدميه وأخذ يتمعَّن فيما قرأه، ثم اخترع التَّفاعُل النَّووي المُتسلسل المُستحث بالنيوترونات.[7] في أقلَّ من أربعٍ وعشرين ساعة، تحوَّلت مسألة تحرير الطاقة من النَّواة من حُكم المُستحيل إلى أنَّها قد حُلَّت من حيث المبدأ. وفي العام اللاحق، سجَّل ليو سيلارد براءة اختراعٍ سريةً لمُفاعلٍ نووي. وفي عام ١٩٣٩، سُجِّلت أول براءة اختراعٍ لسلاحٍ نووي في فرنسا.

المغزى من هذه القصة هو أنَّ المُراهنة على عدم براعة العقل البشري هو رهان خاسر ومُتهور، خصوصًا عندما يكون مُستقبل جنسنا على المحك. مؤخرًا، بدأت موجة إنكارٍ بالظُّهُور داخل مجتمع الذكاء الاصطناعي ذاته، حتى إنها وصلت إلى حدٍّ إنكار احتمالية إحراز أيِّ نجاحٍ فيما يتعلق بأهداف الذكاء الاصطناعي الطويلة الأمد. تخيَّل الأمر كسائقٍ يقود حافلةً وركابها هم البشرية جمعاء، ثم قال السَّائق: «سأقود بكم بأقصى سرعةٍ باتِّجاه جرفٍ صخري، ولكن ثقوا بي؛ سينفد منَّا الوقود قبل أن نصل إلى الحافة!»

أنا لا أزعُمُ بقولي هذا أنَّنا «حتمًا» سننجح في مجال الذكاء الاصطناعي، وأظنُّ أنَّ هذا النَّجاح لو حدث، فلن يكون خلال السنوات القليلة المُقبلة. ومع ذلك، فمن الحكمة والحصافة أن نستعدَّ لهذه الاحتمالية. فإن حدثت، ستكون إيذانًا بعصرٍ ذهبي للبشرية. غير أنَّنا يجب أن نعي حقيقة أنَّنا نُخطِّط لابتكار كياناتٍ تفوق البشر ذكاءً. والسؤال هنا هو: كيف نضمن ألَّا تُسيطر علينا تلك الكيانات؟

ولنأخذ فكرةً عن ضراوة النَّار التي نلعب بها، لننظُر إلى كيفية عمل خوارزميات انتقاء المحتوى في مواقع التَّواصُل الاجتماعي. إن تلك الخوارزميات ليست ذكيةً بوجهٍ خاص، ولكنَّها في موقعٍ تستطيع منه التأثير على العالم أجمع؛ فهي تتحكَّم تحكُّمًا مباشرًا في مليارات البشر. عادةً ما تُصمَّم مثل تلك الخوارزميات لزيادة «مُعدَّل النَّقْر»؛ والذي يعني احتمالية نقر المُستخدم على الأشياء المعروضة أمامه. إذن فأنت تظنُّ أنَّ الحلَّ ببساطةٍ هو أن تعرض الأشياء التي يميل المستخدمون إلى النقر عليها، أليس كذلك؟ هذا غير صحيح. الحلُّ هو أن نُغيِّر من تفضيلات المستخدمين لكي تُصبح اختياراتهم أكثر قابلية للتوقُّع. فالمُستخدم الذي اختياراته أكثر قابلية للتوقع يمكن أن تظهر له المُنتجات التي من المُحتمل أن ينقر عليها؛ ومن ثمَّ تُحقق المزيد من الأرباح. فمثلًا، الأشخاص ذوو الآراء السياسية المُتطرِّفة يتَّسمُون بأنَّ المُنتجات التي يميلون إلى النقر عليها أكثر توقُّعًا. (من المُحتمل أن تكون هناك أيضًا فئة من السِّلع التي يميل الأشخاص الذين يتمسَّكون بآراء سياسيةٍ

مُعتدلةٍ إلى النقر عليها، ولكن ليس من السَّهل تخيُّل ما الذي تنطوي عليه هذه الفئة.) مثلُها كمثل أي كيانٍ منطقي، تتعلم الخوارزمية كيف تُغيِّر من حالة بيئتها — التي هي هنا تفكير المُستخدم — لكي تزيد من الأرباح التي تحصل عليها.[8] وعواقب هذا الأمر تتضمَّن انتشار الفاشية من جديد وفسخ العقد الاجتماعي الذي هو الأساسُ الدَّاعمُ للأنظمة الديموقراطية حول العالم، وربُّما شهدنا حينها نهاية الاتحاد الأوروبي ومُنظَّمة حلْف شمال الأطلنطي. أظنُّ أنَّ هذا ليس بالأمر السيئ بالنسبة لعدة أسطُرٍ برمجيَّة، حتى ولو كان يُساعدها بعض البشر. والآن تخيل معي ما الذي قد تُحدثُه خوارزميات ذاتُ ذكاءٍ حقيقي!

(٣) ما الذي أخطأنا فيه؟

كان الشعار المُحفِّز لأرباب الذكاء الاصطناعي على مرِّ تاريخه هو «كلَّما كانت الآلات أذكى، كان ذلك أفضل.» وفي الحقيقة أنا على قناعةٍ أنَّ هذا القول قول خاطئ؛ لا لأنِّي أشعر بخوفٍ مُبهمٍ من أن الآلات ستحلُّ محلَّنا، بل أراه قولًا خاطئًا بسبب طريقة فهْمنا لماهية الذكاء.

يُعتبر الذَّكاء سببًا رئيسيًّا لما نحن عليه كبشر، ولهذا نُسمِّي أنفسنا «هومو سابيينز»؛ أو «الإنسان العاقل». وبعد ما يزيد عن ألفي عامٍ من التَّفكُّر والتأمُّل في طبيعتنا البشرية، توصَّلنا إلى توصيفٍ للذكاء يُمكن أن يُلخَّص في السَّطر التالي:

> نحن البشر أذكياء ما دامت فعالنا يُتوقَّع منها أن تُحقِّق غاياتنا.

أما بقيَّة خصائص الذكاء، كالإدراك والتَّفكير والتَّعلُّم والابتكار وغيرها، فيُمكن فهمُها في ضوء مُساهماتها في قُدرتنا على التَّصرف بنجاح. ومنذ بدايات مجال الذكاء الاصطناعي، عُرِّف مفهوم الذكاء في الآلات على نفس النَّحو:

> الآلات ذكية ما دامت فعالها يُتوقَّع منها أن تُحقِّق غاياتها.

ولأن الآلات، على عكس البشر، ليست لها غاياتها الخاصة، فنحن من نُودعها الغايات لتُحقِّقها. بمعنًى آخر؛ نحن نبني آلاتٍ تتوخَّى أمثل الحلول، فنُودع فيها ما نريدها أن تُحقِّقه من أهدافٍ، ثم نُطلقُها.

هذا النَّهْج العام ليس مقتصرًا على مجال الذكاء الاصطناعي وحده، بل نراه يتواتر مُتغلغلًا في أُسس مجتمعنا التقنية والرِّياضية. على سبيل المثال، في مجال نظرية التَّحكم؛ ذاك المجال الذي يُصمِّم نُظُم التَّحكُّم في كل شيءٍ حولنا، بدايةً من طائرات الركاب العملاقة إلى مضخَّات الأنسولين، تكون وظيفة النِّظام أن يُبقي على «دالة التَّكلفة» في أدنى قيمةٍ لها؛ هذه الدالة التي تقيس الانحراف عن سلوكٍ ما مرغُوب فيه. وفي مجال الاقتصاد، تُصمم السياسات والآليات لزيادة «منفعة» الأفراد و«رفاهية» الجماعات و«أرباح» الشركات.[9] وفي مجال أبحاث العمليات الذي يسعى لإيجاد حلولٍ للمشاكل التصنيعية واللوجستية المُعقَّدة، يزيد الحل من «مجموعة المكافآت» المُتوقَّعة بمرور الوقت. وأخيرًا، في علم الإحصاء، تُصمم خوارزميات التَّعلُّم لتقليل قيمة «دالة خسارة» متوقعةٍ؛ تُعرِّف كُلفة الوقوع في أخطاءٍ تنبؤية.

من الجليِّ إذن أنَّ هذا الإطار العام؛ والذي سأُسمِّيه من الآن فصاعدًا «النمُوذج القياسي»، هو إطار واسع الانتشار وذو قوةٍ وفعالية. ولكن للأسف، «نحن لا نبغي آلات ذات ذكاءٍ بهذا التَّوصيف».

في عام ١٩٦٠، لفت نوربرت فينر؛ وهو أستاذ أسطوري بمعهد ماساتشوستس للتقنية وأحد أبرز علماء الرياضيات في منتصف القرن العشرين، الأنظار إلى عيوب هذا النمُوذج القياسي. كان فينر قد اطَّلع لتوِّه على لُعبة الدَّامة التي صمَّمها آرثر صامويل والتي طوَّرت من نفسها حتى فاقت صانعها في المُستوى، فقادته هذه التَّجربة إلى كتابة بحثٍ ذي نظرةٍ مُستقبلية مُستبصرة لكنَّه مع ذلك غير مشهور، تحت عنوان: «بعض العواقب الأخلاقية والتقنية للأتمتة.»[10] وها هي الكيفية التي صاغ بها فكرة البحث الأساسية:

> إذا استعملنا، لبُلُوغ الغايات التي ننشُدُها، وسيطًا آليًّا وكُنا لا نستطيع أن نتدخَّل تدخُّلًا كبيرًا في سير عمليَّاته … فجدير بنا أن نتأكَّد أن الغاية التي جعلنا الآلة تسعى لتحقيقها هي الغاية التي نُريد بلوغها حقًّا.

«الغاية التي جعلنا الآلة تسعى لتحقيقها» هي بالضبط الهدف الذي تسعى الآلات لتحقيقه على نحوٍ أمثل في إطار النمُوذج القياسي. ولو وضعنا هدفًا خاطئًا غير الذي نُريده في آلةٍ ذات ذكاءٍ يفوق ذكاءنا البشري، فبلا شكٍّ ستُحقِّقُ ذاك الهدف الخاطئ، وحينها نكون قد خسرنا. وما ذلك التَّصوُّر الكارثي المذكور آنفًا والذي قد تتسبَّبُ فيه مواقع التَّواصل

الاجتماعي إلا دلالة مُنذرة لما قد نجنيه إذا ما وظَّفنا الهدف الخاطئ على نطاقٍ عالمي باستعمال خوارزميات غير ذكية إلى حدٍّ كبير. في الفصل الخامس، سأفصح لكم عن بعض النتائج الأسوأ والأكثر كارثية.

ما قُلتُه يجب ألَّا يُثير دهشتكم مُطلقًا؛ فعلى مدار آلاف السنين ونحن نعلم علم اليقين المخاطر التي تُحيطُ بنا حين نحقق غاية آمالنا بالكامل. وفي كل قصةٍ من القصص التي يُعطى فيها أحد الأشخاص ثلاث أمنياتٍ، دائمًا ما تُبطل الأمنية الثالثة آثار الأمنيتين السَّابقتين عليها.

باختصار، يبدو أنَّ محاولات خلق ذكاء خارق للآلات لا يُمكن إيقافُها، غير أنَّ النَّجاح في تحقيق هذا المأرب قد يكون سبب هلاك الجنس البشري. لكن الوقت لم يتأخَّر بعد. لذلك علينا أن نعرف ما الذي أخطأنا فيه وأن نسْعى لإصلاحه.

(٤) هل يُمكننا إصلاح الأمر؟

يكمُن لُبُّ المشكلة في التَّعريف الأساسي لماهية الذكاء الاصطناعي. فنحن نقول إن الآلات ذكية ما دامت فعالها يُتوقَّع منها أن تُحقِّق «غاياتها»، ومع ذلك فنحن لا نملك أسلوبًا فعَّالًا وجديرًا بالثِّقة لنضمن من خلاله أنَّ «غاياتها» هي نفسها «غاياتنا».

ماذا لو، بدلًا من أن نُصمِّم الآلات لتحقيق «غاياتها»، نُصرُّ بإلحاحٍ على أن تُصمَّم لتحقيق «غاياتنا»؟ ستكون مثل هذه الآلات، إن استطعنا تصميمها، آلاتٍ «ذكية» و«نافعة» للبشر في الوقت ذاته. فلنُحاول إذن أن نصُوغ التَّعريف كما يلي:

> تكون الآلات «نافعةً» ما دامت «فعالها» يُتوقَّع منها أن تُحقِّق «غاياتنا».

رُبَّما كان هذا هو ما كان يجبُ علينا فعله منذ البداية.

أصعب جُزءٍ بلا ريبٍ هو أنَّ غاياتنا موجودة بداخلنا — أي داخل كلِّ فردٍ من الثَّمانية مليار بشريٍّ بكل ما حُبينا به من تنوُّعٍ واختلافٍ عظيمين — وليس بداخل الآلات. وبالرغم من ذلك، يُمكنُنا أن نبني آلاتٍ نافعة بنفس هذا المعنى. إن هذه الآلات ستكون غير مُتيقنة من ماهية غاياتنا — ولكنَّا في النِّهاية أيضًا على نفس الحال — لكن هذه ميزة لا عيب (أي إنها شيء حسن لا شيء سيِّئ). فعدم اليقين بشأن الغايات يضمنُ أن تظلَّ الآلات بالضرورة مُذعنةً للبشر؛ فلسوف تطلُبُ الإذن، وتتقبَّلُ التَّصحيح، وتستسلم لأوامر إيقاف تشغيلها.

إذا استبعدنا افتراض أنَّ الآلات يجب أن تُلقَّم بغاياتٍ وأهدافٍ مُحدَّدة، حينها سنُضطر إلى هدم جُزءٍ من أسس الذكاء الاصطناعي ثمَّ استبداله؛ وهذا الجزء هو المفاهيم الأساسية لما نُحاول الوصول إليه في هذا المجال. كما يعني هذا أيضًا أن نُعيد بناء جُزءٍ كبيرٍ من البنْية الفوقيَّة؛ وهي تلك الأفكار والأساليب المُتراكمة التي تُشكِّل أساس الذكاء الاصطناعي الحالي. سينتُج عن ذلك علاقة جديدة بين البشر والآلات؛ تلك العلاقة التي أرجو أن تُمكِّننا من اجتياز العُقُود القليلة القادمة بنجاح.

الفصل الثاني

مفهوم الذكاء في البشر والآلات

عندما نصل إلى طريقٍ مسدُودٍ، فمن الحكمة أن نعود أدراجنا ونقتفي آثار سيرنا لنقف على أي طريقٍ خاطئٍ سلكناه. ولقد حاججتُ بأنَّ النمُوذج القياسي للذكاء الاصطناعي ما هو إلا طريق مسدُود؛ ذلك النَّموذج الذي تعكفُ الآلات في ظلِّه على الوُصُول بأفضل الطُّرُق إلى الغايات المُحدَّدة التي أودعها البشر إيَّاها. والمُعضلة هنا ليست أنَّنا قد «نفشل» في بناء نظم الذكاء الاصطناعي، بل في أنَّنا قد «ننجح» نجاحًا عظيمًا. فمفهوم النَّجاح في مجال الذكاء الاصطناعي خاطئ بالكُلِّية.

فهيا بنا إذن نعُد أدراجنا ونقتفِ آثارنا من بداية الطَّريق. لنُحاول معًا أن نفهم كيف تبلور مفهُوم الذكاء لدينا وكيف طُبِّق على الآلات. حينها سنحظى بفُرصةٍ لنقترح مفهومًا أفضل لما يُمكن أن يُعدَّ كنظام ذكاءٍ اصطناعي جيد.

(١) الذَّكاء

ما نواميسُ هذا الكون؟ وكيف بدأت الحياة؟ وأين هي سلسلة مفاتيحي؟ تلك أسئلة جوهرية جديرة بالتأمُّل والتَّفكير. ولكن من عساه يسأل مثل هذه الأسئلة؟ وكيف سأجيبُ عنها؟ وكيف لحفنةٍ من الخلايا؛ تلك الكُرة ذات اللون الوردي المائل للرمادي التي تُشبه المُهلَّبية والتي نُسميها الدِّماغ، أن تُدرك وتفهم وتتنبَّأ وتتدبَّر بدهاءٍ أمر عالمٍ من الفضاء الشَّاسع والفسيح؟ ثم بدأ العقل يسبرُ أغوار نفسه.

مُنذُ آلاف السِّنين ونحن نسعى لفهم كيف تعمل عقولنا. في البداية، كان الفُضُول هو ما يدفعُنا إلى ذلك، بجانب مساعي الإدارة الذاتية، وتحصيل القُدرة على الإقناع، ولهدفٍ عمليٍّ آخر وهو تحليل البراهين الرياضية. ومع ذلك، فكلُّ خُطوةٍ نخطُوها إلى

الأمام في طريق فهمنا لآلية عمل العقل، هي في الوقت ذاته خُطوة تُقرِّبُنا من مُحاكاة القدرات العقلية في آلةٍ من صُنع الإنسان؛ والتي بدورها خُطوة إلى الأمام في مجال الذكاء الاصطناعي.

إن فهمنا لماهيَّة الذكاء سيُساعدنا في فهم كيف نبنيه في آلات. ولن نتوصَّل إلى هذا الفهم من خلال اختبارات معدَّل الذكاء ولا حتى في اختبارات تورينج، بل هو يقبعُ في علاقةٍ بسيطةٍ بين ما ندركه وما نُريده وما نفعله. يُمكن القول إن أي كيانٍ يُعدُّ ذكيًّا ما دامت فعالُه يُتوقَّع منها أن تُحقِّق ما يريده، مع الأخذ في الاعتبار ما يدركه.

(١-١) الأصول التَّطوريَّة

تأمَّلْ إحدى الجراثيم البسيطة مثل الإي كولاي (جرثومة المعدة). ستجدها مُزوَّدةً بنحو نصف دزينة من الأسواط؛ وهي مجسات طويلة ورقيقة كالشَّعرة تدور قواعدها إما في اتجاه عقارب الساعة أو عكسه. (أمَّا المُحرِّك الدَّوار ذاته فهو آية عظيمة، ولكن ليس هذا مقام الحديث عنه.) وبينما تطفو هذه الجرثومة في بيئتها السائلة؛ الجزء الأسفل من جهازك الهضمي، تُبادل بين تدوير أسواطها في اتجاه عقارب الساعة ممَّا يجعلها تتقلَّب في مكانها، وبين تدويرها في عكس اتجاه عقارب الساعة، فتصير الأسواط كحبلٍ مجدُولٍ يُشبه مروحةً دافعةً مما يُمكِّنُ الجرثومة من السِّباحة في خطٍّ مُستقيم. وهكذا، فإنَّ هذه الجُرثُومة تقوم بنوع من التحرُّك العشوائي؛ تسبح ثمَّ تتقلَّب، ثمَّ تسبح ثم تتقلَّب، وهذا يُتيح لها العثور على جزيئات الجُلوكوز وامتصاصها بدلًا من البقاء ساكنةً مكانها والموت جوعًا.

لو كانت هذه هي الحكاية برُمَّتها، لم نكن لنقول إن جُرثُومة الإي كولاي ذكية على وجه الخصوص؛ لأنَّ فعالها لا تعتمد على أيِّ نحوٍ على البيئة المحيطة؛ فهي بهذه الصُّورة لا تتَّخذ أي قراراتٍ، بل تؤدي سُلُوكًا ثابتًا بناه التَّطور في جيناتها. ولكن ليست القصة كاملةً. فعندما تستشعر هذه الجُرثُومة ازديادًا في تركيز الجُلوكُوز، تبدأ في السِّباحة لمسافةٍ أطول وتُقلِّل الالتفاف، والعكسُ صحيح عندما تستشعر نقصًا في تركيز الجُلوكوز. فما تفعله هذه الجُرثومة إذن (السِّباحة صوب جزيئات الجُلوكُوز) يُتوقَّع منه على الأرجح أن يُحقِّق ما تريده (لنفرض أن ما تريده هو امتصاص المزيد من الجُلوكوز) بناءً على ما أدركته (ازدياد تركيز الجُلوكوز).

رُبَّما تُفكِّر وتقول: «ولكن ألم يدمج التَّطور هذا التَّصرف في جيناتها أيضًا؟! كيف لها إذن أن تُعدَّ كيانًا ذكيًّا؟» أقول لك إن هذا خطٌّ تفكيرٍ شديد الخطورة؛ فالتَّطوُّر هو

من دمج التصميم الأساسي لدماغك في جيناتك أيضًا، ولا أظنُّ أنَّك سترغبُ في نفي صفة الذَّكاء عنك بناءً على هذا الاعتقاد. ما أرمي إليه هو أنَّ ما دمجه التَّطوُّر في جينات جُرثومة الإي كولاي، الذي هو نفسه ما فعله في جيناتك أنت، هو مُجرَّد آلية يتغيَّر بموجبها سُلوك الجُرثومة طبقًا لما تُدْركه في بيئتها المُحيطة. فالتَّطور لا يعلم مُسبقًا أين سيكون موقع جزيئات الجُلوكوز أو أين هي سلسلة مفاتيحك، لذلك فغرسُ القُدرة التي تُؤهِّلُك للعُثُور عليها هو ثاني أفضل الخيارات.

إن هذه الجرثومة ليست شديدة الذكاء. فعلى حدِّ معرفتنا، هي لا تتذكَّر الأماكن التي مرَّت بها؛ فإذا تحرَّكت من النقطة «أ» إلى النقطة «ب» ولم تجد جزيئات الجُلوكُوز، فمن المُحتمل أن تعود إلى النقطة «أ» مرةً أخرى. وإذا هيَّئنا بيئةً ما حيث تقُود جزيئات مُدرجة من الجُلوكوز المُغري إلى نُقطةٍ من الفينُول الذي يُعتبر سُمًّا للجُرثومة، ستظلُّ تتبع جزيئات الجُلوكوز المُؤدِّية إلى السُّم. ولن تتعلم أبدًا؛ فلا دماغ لدَيها؛ فما لدَيها هو مجرد بعض التفاعلات الكيميائية البسيطة التي تُساعدها في القيام بمهامها.

ثمَّ حدثت خطوة كبيرة للأمام مع ظهور «جُهد الفعل»؛ وجُهدُ الفعل هذا هو نوع من الإشارات الكهربية التي ظهرت لأول مرة في الكائنات الوحيدة الخلية قبل ما يُقارب المليار سنة. ثمَّ طوَّرت الكائنات المُتعدِّدة الخلايا فيما بعدُ خلايا مُتخصِّصة تُسمَّى «العصبُونات» والتي تَستخدم جهد الفعل الكهربي لنقل الإشارات داخل الكائن الحي بسُرعةٍ فائقةٍ؛ تصل إلى ١٢٠ مترًا في الثانية أو ٢٧٠ ميلًا في الساعة. وتُسمَّى الروابط بين العصبُونات بـ «المشابك العصبية». تُحدِّد قوة هذه المشابك العصبية حجم الإثارة الكهربية التي تنتقل من عصبُونٍ إلى آخر، وبتغيير قوة هذه المشابك العصبية يحصُل التَّعلُّم لدى الحيوانات.[1] إن التَّعلُّم يمنحُ مزيَّةً تطوُّريةً هائلة؛ فمن خلاله تستطيع الحيوانات التأقلُم والتَّعايش مع مجموعةٍ هائلة من الظُّروف، كما يُسرِّع من وتيرة التَّطوُّر ذاتها.

في البداية، رُتِّبت العصبُونات في «شبكاتٍ عصبيةٍ» مُوزَّعة في جسد الكائن الحي لتُساعد في تنظيم أنشطة مثل الأكل والهضم، أو تنظيم الانقباضات الموقُوتة لخلايا العضلات على نطاقٍ كبير. وما نراه من حركةٍ رشيقةٍ لقناديل البحر ما هي إلا نتيجة لشبكة عصبية؛ فليس لقناديل البحر دماغ إطلاقًا.

أما الأدمغة فقد ظهرت فيما بعد، جنبًا إلى جنبٍ مع أعضاء الحسِّ المُعقَّدة كالأعيُن والآذان. فبعد ظهور قناديل البحر ذات الشبكات العصبية بمئات الملايين من الأعوام، وُجدنا نحن البشر بأدمغتنا الضَّخمة؛ مائة مليار عصبُون (١٠^١١) وكوادريليون مشبكٍ

عصبي (١٠^١٥). ورغم أن الدِّماغ البشري بطيء بالمقارنة بالدَّوائر الإلكترونية؛ فإنَّه يُعتبر سريعًا إذا ما قُورن بمُعظم العمليات الحيوية؛ فزمن الدَّورة الكهربية لكُلِّ تغيير حالة يُقدَّر ببضعة ميلِّي ثانية. وعادةً ما يصف البشر دماغهم بأنَّه «أكثر الأشياء تعقيدًا في الكون»، ومع أنَّ هذا الادِّعاء قد لا يكون صحيحًا، فإنَّه عُذر مقبول نُقدِّمُه حين تُذكر حقيقة أنَّ فهمنا لآلية عمله ما يزال ضئيلًا. وفي حين أنَّنا نعرف قدرًا عظيمًا عن الكيمياء الحيوية للعصبُونات والمشابك العصبية، وكذلك عن البُنى التَّشريحية للدماغ، فإنَّ العمليات العصبية التي تحدُث على المستوى «المعرفي» — كالتَّعلُّم والإدراك والتَّذكُّر والتفكير والتَّخطيط واتِّخاذ القرارات وهلُمَّ جرًّا — ما تزال غير معروفة.[2] (ربَّما سيتبدَّل الحال عندما يزداد فهمُنا للذكاء الاصطناعي، أو عندما نُطوِّر أدواتٍ أدق لقياس نشاط الدِّماغ.) لذلك عندما يقرأ المرء في الإعلام أنَّ إحدى تقنيات الذكاء الاصطناعي «تُضاهي الدِّماغ البشريَّ في آلية عملها»، لا يعرف هل هذا الكلام هو مجرَّد افتراضٍ أم محْض خيال.

أما بالنسبة لمجال «الوعي»، فنحن لا نعرف عنه شيئًا، لذلك لن أكتب عنه حرفًا. فلا أحد في مجال الذكاء الاصطناعي يسعى لبناء آلات ذات وعيٍ، ولا أحد يعرف من أين يبدأ إن كان يسعى لذلك، ولا يُوجَد أي سُلوك يتطلَّب وعيًا كمُتطلَّبٍ أساسي له. لنفترض أنِّي أعطيتُك برنامجًا ثمَّ سألتك: «هل يُمثِّل هذا البرنامج تهديدًا للبشرية؟» ستفحصُ شفرة البرنامج وتُحلِّلها وبالفعل عند تشغيلها، تجد أنها تبدأ في صياغة وتنفيذ خطةٍ نتاجها في النِّهاية سيكون هلاك الجنس البشري، تمامًا كما يصُوغ ويُنفِّذ برنامج خاص بلعب الشطرنج خُطَّةً لهزيمة أي لاعبٍ بشري يُنازله. والآن لنفترض أنِّي قُلتُ لك إن هذه الشَّفرة ستُنشئ عند تشغيلها ضربًا من ضُرُوب الوعي في الآلات، هل سيُؤثِّر هذا على توقُّعاتك؟ لا، إطلاقًا. فلا شيء سيتغيَّر ألبتَّة.[3] فتوقُّعاتك لسُلُوك البرنامج ستظلُّ كما هي، وهذا لأنك قد بنيت تلك التَّوقُّعات على ما رأيته من شفرة. فكل ما نراه من حبْكاتٍ لأفلام هُوليود حول آلاتٍ تُصبح ذات وعيٍ على نحوٍ مُبهمٍ ويُعادُون البشر ويكرهونهم لأسبابٍ غامضةٍ، كلُّ هذه الحبْكات تُسيء فهْم الأمر؛ فالمُهمُّ هو الكفاءة لا الوعي.

من أهمِّ الجوانب المعرفية للدماغ التي بدأنا نفهمها ما يُعرف باسم «نظام المُكافأة». وهذا النِّظام هو نظامُ إشارةٍ داخلي يربط ما بين السُّلُوك والمُحفِّزات الإيجابية أو السَّلبية عن طريق مادة الدُّوبامين. وقد اكتُشفت آلية عمل هذا النِّظام في أواخر خمسينيات القرن

الماضي على يد عالِم الأعصاب السويدي نيلس-آكي هيلارب ومُعاونيه. إن هذا النِّظام يدفعنا إلى السَّعي وراء المُحفزات الإيجابية كالطعام الحُلو المذاق الذي يزيد من إفراز مادة الدوبامين، ويحُثُّنا على تجنُّب المُحفِّزات السَّلبية كالجُوع والألم التي تُنقص من مُعدَّلات تلك المادة. وإذا نظرنا إلى هذا النظام، سنجده يُشبه إلى حدٍّ ما آلية السَّعي وراء جزيئات الجُلوكُوز عند جُرثومة الإي كولاي، ولكن على مستوى أعقد بكثير. فهذا النِّظام مُصمَّم بأساليب للتَّعلُّم بحيث يصير سلوكنا بمرور الوقت أكثر فعاليةً في الحصول على الإثابة. كما يُتيح لنا أيضًا خاصية اللذة المُؤجَّلة؛ فنتعلَّم كيف نشتهي الأشياء كالمال مثلًا، الذي سيمنحُنا إثابة لاحقةً مُحتملة بدلًا عن إثابةٍ فورية. وأحد الأسباب الكامنة وراء فهمنا لنظام المُكافأة في الدماغ هو أنَّه يُشابه أسلوب «التَّعلُّم المُعزَّز» الذي طُوِّر في أروقة مجال الذكاء الاصطناعي والذي نملك حوله نظريةً مُثبتةً ومُحكمةً.[4]

من وجهة نظرٍ تطوُّريَّة، يُمكننا اعتبار نظام المُكافأة في الدماغ، مثله كمثل آلية السَّعي وراء جزيئات الجُلوكُوز عند جُرثُومة الإي كولاي، بمنزلة طريقةٍ لتحسين الصَّلاحية التَّطورية. فالكائنات ذات الآليات الأكثر فعاليَّةً في السَّعي وراء المُكافأة — كالعُثُور على طعامٍ لذيذٍ، وتجنُّب الشعور بالألم، ومُمارسة النَّشاط الجنسي، وما إلى ذلك — يحظون بفُرصٍ أكثر لنقل جيناتهم للأجيال اللاحِقة. من الصعب جدًّا على أيِّ كائنٍ من الكائنات الحية أن يُحدِّد ماهيَّة التَّصرُّفات التي قد تصل به على المدى الطويل إلى أن ينقل جيناته للأجيال اللاحقة بنجاح، لذلك سهَّل التَّطور هذا الأمر لنا بأن زوَّدنا بعلاماتٍ إرشاديةٍ داخلية على طول الطَّريق.

ومع ذلك، تلك العلامات الإرشادية ليست مثاليَّة. فهناك طرق للحصول على الإثابة والتي رُبما «تُقلِّل» من احتمالية أن ينقل الفرد جيناته إلى أجيالٍ قادمة. على سبيل المثال، تعاطي المُخدِّرات، والإفراط في تناول المشروبات الغازية المسكرة، والانهماك في ألعاب الفيديو لمدة ثماني عشرة ساعةً مُتواصلة يوميًّا؛ كل هذه الأفعال تأتي بنتائج عكسية فيما يتعلَّق بعملية التَّناسُل والتَّوارث. بالإضافة إلى ذلك، إنك إذا أُعطيت تحكُّمًا كهربيًّا مباشرًا في نظام المُكافأة في جسدك، فعلى الأرجح أنَّك ستظلُّ تُحفِّز النظام ذاتيًّا دون توقُّفٍ حتى تلقى حتفك.[5]

إن اختلال نظام المُكافأة والصَّلاحية التَّطورية لا يؤثر على البشر فحسب. فعلى سبيل المثال، على جزيرةٍ صغيرةٍ قُبالة الشَّواطئ البنميَّة يعيش حيوان الكسلان القزم الثلاثي

أصابع القدم، والذي اتَّضح أنَّه يُدمن مادةً تُشبه في تأثيرها عقارًا مُهدِّئًا يُسمَّى الفلْيُوم من خلال تغذيته على أوراق أشجار المانجروف الحمراء، وأنه قد يكون مُهدَّدًا بالانقراض.[6] من الواضح إذن أنَّ نوعًا بأكمله يُمكن أن يندثر إذا عثر على ظروف بيئية مناسبة حيث يُمكنُه أن يُشبع نظام المُكافأة داخله على نحوٍ فيه سُوء تكيُّف.

مع ذلك، وباستثناء حالات الإخفاق العارضة تلك، فإنَّ تعلُّم كيفية زيادة الحصول على المُكافأة في البيئات الطبيعية عادةً ما سيُحسِّن من فُرص الفرد في نقل جيناته، ومن فُرص بقائه في ظل التَّغيُّرات البيئية.

(١-٢) تسارُع التَّطوُّر

التَّعلُّم مُفيد لأسباب غير البقاء والتَّكاثُر؛ فهو يُسرِّع أيضًا من وتيرة التَّطوُّر. كيف يُمكن هذا؟ ففي نهاية المطاف، التَّعلُّم لا يُغيِّر من حمضنا النَّووي، أما التَّطوُّر فما هو إلا تغيير الحمض النَّووي على مدار أجيالٍ مُتعاقبة. لقد طُرحت العلاقة بين التَّطوُّر والتَّعلُّم عام ١٨٩٦ على يد عالم النَّفس الأمريكي جيمس بالدوين،[7] كما طرحه قبل ذلك عالم السُّلوك الحيواني البريطاني كونوي لويد مورجان[8] ولكن أطروحته لم تُقبل بوجهٍ عامٍّ في ذلك الوقت.

يُمكن فهم «ظاهرة بالدوين»، كما تُسمَّى الآن، بتخيُّل أنَّ التَّطوُّر مُخيَّر؛ إما أن يبني كائناتٍ «غريزيَّة» تكون كُلُّ رُدُود أفعالها مُدمجة فيها مُسبقًا، أو أن يبني كائناتٍ «قادرة على التأقلُم» تتعلَّم ما الذي يجب عليها فعلُه. ولهدف إيضاح الأمر أكثر، تخيَّل معي أنَّ الكائن «الغريزي» المثالي يُمكن أن يُشفَّر برقمٍ من ستِّ خاناتٍ، وليكن مثلًا: ٤٧٢١١٦، بينما في حالة الكائن «القادر على التأقلم» يُحدِّد التَّطوُّر له ثلاث خاناتٍ فقط: ٤٧٢***، وعلى الكائن أن يُكمل باقي الشَّفرة من خلال ما يتعلَّمه في مسيرة حياته. إذن فمن الواضح أنَّ التَّطوُّر إذا كان عليه أن يُحدِّد الخانات الثَّلاث الأُوَل فقط من الشَّفرة، فمهمته ستكون أسهل بكثير؛ فالكائن «القادر على التأقلم» إذْ يكتشف أرقام الخانات الثَّلاث الأخيرة، يُنجز في حياةٍ واحدةٍ ما قد يستغرق التَّطوُّر عدة أجيالٍ ليُنجزه. وهكذا، وبفرض أنَّ الكائنات القادرة على التأقلُم يُمكنُها البقاء خلال رحلة التَّعلُّم، يبدو أنَّ القُدرة على التَّعلُّم تُمثِّل طريقًا تطوُّريًّا مُختصرًا. وتُشير تجارب المُحاكاة الحوسبيَّة إلى أنَّ ظاهرة بالدوين هي ظاهرة حقيقية.[9] ويقتصر تأثير الثَّقافة على تسريع العملية؛ وهذا لأنَّ الحضارة المنظمة

دائمًا ما تحمي الفرد أثناء عملية تعلُّمه وتنقلُ له المعلومات التي قد يحتاج إلى تعلُّمها بنفسه من جديدٍ إن لم تُنقل له.

أما ظاهرة بالدوين، فقصَّتُها مُشوِّقة لكنَّها ناقصة؛ فهي تفترضُ أنَّ التَّعلُّم والتَّطوُّر يمضيان معًا بالضرورة في اتجاهٍ واحد. ومن ذلك المُنطلق، فهي تفترضُ أنَّ أي إشارةٍ لاستجابةٍ داخليةٍ تُحدِّد اتِّجاه عملية التَّعلُّم داخل الكائن تتَّفق اتفاقًا وثيقًا مع الصَّلاحية التَّطورية. ولكن كما رأينا في حالة حيوان الكسلان القزم الثُّلاثي أصابع القدم، فإن مثل هذا الافتراض يبدو أنَّه خاطئ. ففي أفضل الأحوال، لا تُمدُّ آليات التَّعلُّم المُدمجة الكائن سوى بتلميحاتٍ أوَّليةٍ عن العواقب الطويلة الأمد لأيِّ فعلٍ بالنِّسبة إلى الصَّلاحية التَّطورية. من ناحيةٍ أخرى، علينا أن نسأل: «كيف تسنَّى لنظام المُكافأة أن يُوجَد في الكائنات في المقام الأول؟» والإجابة قطعًا هي أنَّه وُجِد عن طريق عمليةٍ تطوُّريَّةٍ تحوي بداخلها آلية استجابةٍ تتوافق على الأقل بعض الشيء مع الصَّلاحية التَّطورية.[10] من الجليِّ أن آلية التَّعلُّم التي تحثُّ الكائنات على النفور من الرِّفاق المُحتمَلين، وتدفعُهم في الوقت ذاته إلى التَّقرُّب من المُفترسين لن تدُوم طويلًا.

وهكذا، فالشُّكرُ موصُول إلى ظاهرة بالدوين على إيضاح حقيقة أنَّ العصبُونات بقدرتها على التَّعلم وحلِّ المُشكلات، تنتشر انتشارًا واسعًا في مملكة الحيوان. وفي الوقت ذاته، من المُهم لنا أن نعي أنَّ التَّطوُّر لا يَعنيه حقًّا إن كنت كائنًا ذا دماغٍ أو تُعمل عقلك بأفكارٍ مُدهشة. فما أنت إلا مُجرَّد «كيان» بالنِّسبة إليه؛ أي ما أنت إلا شيء ما يفعل الفعل. ورُبَّما تكون الصِّفات العقلية القيِّمة؛ كالتَّفكير المنطقي والتَّخطيط المُتأنِّي والحكمة والفطنة والخيال والإبداع، أساسيةً في تكوين كيان ذكي، ورُبما كانت غير أساسية. وأحد الأسباب التي تُضفي على مجال الذكاء الاصطناعي سحرًا وجاذبيةً هو أنَّه يُقدِّم مُقترحًا لفهم هذه القضايا؛ مُقترحًا قد يُوصلُنا إلى فهمٍ لكيف تُتيح تلك الصِّفات العقلية تكوين سُلُوك ذكي، ولماذا من المُستحيل أن نُصدر سُلُوكًا ذكيًّا حقيقيًّا دونها.

(١-٣) عقلانيَّة الفرد

مُنذُ بدايات الفلسفة الإغريقيَّة القديمة، انحصر مفهوم الذكاء في القدرة على الاستيعاب وإعمال الفكر والتَّصرف «بفعالية».[11] وعلى مرِّ القُرُون، أخذ هذا المفهوم يتوسَّع في قابليته للتَّطبيق، كما أصبح أكثر تحديدًا في تعريفه.

كان أرسطو أحد الذين بحثوا في مفهوم التفكير الفعال؛ وهي طُرُق الاستدلال المنطقي التي تُفضي إلى نتائج صحيحة بناءً على مقدمات صحيحة. كما بحث أيضًا عمليَّة اتِّخاذ قرارات الأفعال، والتي تُسمَّى أحيانًا بـ «التفكير العملي»، ثمَّ اقترح أنَّ هذه العمليَّة تنطوي على الاستدلال بأنَّ مسارًا ما سيُحقِّق هدفًا منشُودًا ما:

> نحن لا نتفكَّر في الغايات، بل نتدبَّر الوسائل التي تُوصِّلنا إليها. فالطَّبيب لا يُفكِّر إن كان سيشفي مريضه أم لا، والخطيب الواعظ لا يُفكِّر إن كان سيُقنع مُستمعه أم لا. … بل يفترض كلاهُما الغاية المرجُوَّة، ثمَّ يدرُسان بتروٍّ كيف يصلان إلى تلكُما الغاية وأي السُّبُل يسلكان، ثُمَّ يقفان على مقدار سُهُولة تلك السُّبُل ومدى فعاليَّتها وكفايتها؛ وإذا تراءى لهُما أنَّ الغاية لا تُدرَك إلا بسبيلٍ واحدٍ لا غير، حينئذٍ يتأمَّلان «كيف» سيُدركانها بهذا السَّبيل، بل وكيف سيظفران بهذا السَّبيل، وهكذا إلى أن يصلا إلى العلَّة الأولى … وما يأتي أخيرًا في سلسلة التَّحليل، يأتي أولًا في ترتيب الوجود. وإذا ما تأكَّدنا أنَّ الغاية بعيدة المنال، ضجرنا بالبحث وتركناه؛ ومثال ذلك، متى كُنا نحتاجُ إلى المال ولا نستطيع أن نُصيبه؛ غير أنَّه إذا بدا أنَّ غايةً ما مُمكنة الحُدُوث، فإنَّنا نبذُل الجُهد لنيلها.[12]

يحقُّ للمرء أن يقول إن هذه الفقرة قد أرست أسُس الفكر الغربي حول العقلانيَّة منذ ما يربُو على الألفي عام. فهي تُخبرنا أنَّ «الغاية»، وهي مُراد الإنسان، تكون مُفترضة وثابتة. كما تُخبرنا أيضًا أنَّ التَّصرُّف العقلاني هو التَّصرُّف الذي يصل بصاحبه إلى الغاية المُرادة «بسُهُولةٍ وكفاءةٍ» استنادًا إلى الاستنتاج المنطقي عبر سلسلةٍ من الأفعال.

يبدُو طرح أرسطُو هذا طرحًا معقولًا، لكنَّه لا يُقدِّم تفسيرًا شاملًا للسُّلوك العقلاني. وتحديدًا، فإنَّه يغفل عن مُشكلة الارتياب وعدم اليقين. ففي العالم الحقيقي، يميل الواقع إلى التَّدخُّل، وقليل من الأفعال أو سلاسل الأفعال هي التي تضمن حقًّا تحقيق غاياتك المنشُودة. على سبيل المثال، أنا أكتبُ هذه الجُملة التي تقرءونها في يوم أحدٍ مُمطر في مدينة باريس، وفي يوم الثُّلاثاء تُقلع طائرتي المُتوجِّهة إلى مدينة روما في الساعة الثَّانية والرُّبع عصرًا من مطار شارل ديجول الذي يبعُد حوالي خمسة وأربعين دقيقة من بيتي. خُطَّتي هي أن أُغادر مُتَّجهًا إلى المطار حوالي الساعة الحادية عشرة والنِّصف ظُهرًا مما يمنحُني

مُتسعًا من الوقت، ولكن قد يعني هذا أني قد أجلسُ قُرابة الساعة على الأقل مُنتظرًا في صالة المُغادرة. هل أنا هكذا «مُتأكِّد» من أنِّي سألحق بالطائرة؟ قطعًا لا. فلرُبَّما واجهتُ ازدحامًا مروريًّا خانقًا، أو يُعلن سائقو سيارات الأجرة الإضراب، أو رُبَّما تتعطَّل سيارة الأجرة التي أستقلُّها أو يُقبض على السائق بعد مُطاردةٍ بسبب السُّرعة القُصوى، وهلُمَّ جرًّا. ولأتجنَّب كُلَّ ذلك، عليَّ إذن أن أتَّجه إلى المطار يوم الاثنين؛ يوم كامل مُقدَّمًا. بلا شك سيُقلِّلُ هذا التَّصرُّف كثيرًا من احتمالات عدم اللحاق برحلتي، ولكن تخيُّل قضاء ليلةٍ في صالة المُغادرة لا يبدو مشهدًا جيدًا أبدًا. بمعنًى آخر، تتضمَّن خُطَّتي «مُقايضةً» بين حتميَّة النَّجاح وكُلفة ضمان مثل هذه الحتميَّة. الخُطَّة التالية لشراء منزلٍ تتضمَّن أيضًا عملية مُقايضةٍ مُماثلة؛ تشتري بطاقة يانصيب، فتربح مليون دولار ثمَّ تشتري المنزل. إن هذه الخُطَّة تصل بصاحبها إلى الغاية المُرادة «بسهولةٍ وكفاءةٍ»، ولكن تقلُّ كثيرًا احتمالات أن تنجح. الفرقُ بين تلك الخُطَّة الطائشة لشراء منزلٍ وخُطَّتي الأوقع والأكثر حصافةً للذهاب إلى المطار يكمنُ في احتمالية الحُدُوث. فكلتا الخُطَّتين فيهما مُقامرة ومُجازفة، ولكن إحداهُما تبدو أكثر عقلانيَّةً من الأخرى.

وهنا يتَّضح أنَّ المقامرة كان لها دورٌ رئيسي في تعميم طرح أرسطُو لتُعلِّل مُشكلة عدم اليقين. في العقد السادس من القرن السادس عشر، طوَّر عالم الرِّياضيات الإيطالي جيرولامو كاردانو أوَّل نظريةٍ دقيقةٍ رياضيًّا للاحتمال؛ وذلك باستخدام ألعاب النَّرد كمثالٍ رئيسي. (ولكن مع الأسف لم تُنشَر أبحاثُه إلا عام ١٦٦٣.)[13] وفي القرن السابع عشر، بدأ المُفكِّرون الفرنسيُّون، بما فيهم أنطوان أرنولد وبليز باسكال، في البحث عن جوابٍ لمسألة القرارات العقلانية في المقامرة،[14] وقد كان ذلك لأسبابٍ رياضيَّةٍ بحتة. تأمَّل معي الرِّهانين التاليين:

(أ) احتماليَّة ٢٠ بالمائة أن تربح ١٠ دولارات.

(ب) احتماليَّة ٥ بالمائة أن تربح ١٠٠ دولار.

قد تُشابه الأطرُوحة التي عرضها عُلماء الرِّياضيات ما تجُود به قريحتُك في هذه المسألة؛ وهي أن نُقارن «القيمة المُتوقَّعة» لكُلٍّ من الرهانين، أي مُتوسِّط المبلغ الذي قد تحصُل عليه من كُلِّ رهان. فالقيمة المُتوقَّعة للرهان «أ» هي ٢٠ بالمائة من العشرة دولارات؛ أي دولاران. أما الرهان «ب»، فقيمته المُتوقَّعة هي ٥ بالمائة من المائة دولار؛ أي خمسة دولارات. لذلك، وطبقًا لهذه الأطرُوحة، نجدُ أنَّ الرَّهان «ب» هو الأفضل. وعليه يُمكن

القول إنها أطرُوحة منطقية، وهذا لأنَّنا إذا قامرنا بنفس الرَّهان مرارًا وتكرارًا، فالمُقامر الذي سيتبع القاعدة سينتهي به المطافُ وقد ربح أموالًا أكثر ممَّن لمْ يتبعها.

في القرن الثَّامن عشر، لاحظ عالم الرِّياضيات السويسري دانييل برنولي أنَّ هذه القاعدة يبدو أنها لا تنطبق على المبالغ الكبيرة من الأموال.[15] فعلى سبيل المثال، تأمَّل معي الرهانين التاليين:

(أ) احتماليَّة ١٠٠ بالمائة أن تربح ١٠٠٠٠٠٠٠ دولار. (القيمة المُتوقَّعة هي ١٠٠٠٠٠٠٠ دولار.)

(ب) احتماليَّة ١ بالمائة أن تربح ١٠٠٠٠٠٠١٠٠ دولارٍ. (القيمة المُتوقَّعة هي ١٠٠٠٠٠٠١ دولارٍ.)

السَّواد الأعظم من قُراء هذا الكتاب ومُؤلِّفُه إلى جانبهم، سيُفضِّلون الرهان «أ» على الرهان «ب»، رغم أنَّ قاعدة القيمة المُتوقَّعة تُشير إلى عكس ذلك! وهُنا افترض دانييل برنولي أنَّ الرهانات لا تُقيَّم وفقًا لقيمتها النَّقديَّة المتوقعة، ولكن حسْب «منفعتها» المُتوقَّعة. والمنفعة — وهي صفةُ كون الشَّيء مُفيدًا أو ذا نفعٍ للشخص — هي، كما اقترح دانييل، كمِّية ذاتية داخلية تتعلَّق بالقيمة النَّقديَّة لكنَّها مُختلفة عنها. وتفصيلًا؛ المنفعة تُظهر عوائد مُتناقصة بالنِّسبة إلى الأموال. وهذا يعني أنَّ منفعة أي مقدارٍ من المال لا تتناسب تناسُبًا دقيقًا مع مقداره، لكنَّها تنمُو ببطءٍ أكثر. ومثال ذلك هو أنَّ منفعة ربح ١٠٠٠٠٠٠١٠٠ دولارٍ أقل بما يزيد عن مائة مرة من منفعة ربح ١٠٠٠٠٠٠٠ دولار. السؤال هو: أقل بكم تحديدًا؟ اسأل نفسك! ما هي نسبة الاحتماليَّة المُرضية لك لتُراهن على ربح مليار دولارٍ وتتخلَّى عن عشرة ملايين مضمُونة؟ سألتُ هذا السُّؤال في صفٍّ لطلبة الدِّراسات العُليا، وكانت إجاباتهم تقعُ قُرب نسبة الخمسين بالمائة، وهذ يعني أنَّ الرهان «ب» سيكون ذا قيمةٍ مُتوقَّعةٍ مقدارُها ٥٠٠ مليون دولارٍ؛ وذلك حتى يُماثل جاذبية الرهان «أ». واسمحُوا لِي أن أكرِّر هذه النُّقطة وأقول إن الرهان «ب» سيكون ذا قيمةٍ نقديةٍ مُتوقَّعةٍ أعلى بخمسين مرَّةً من الرهان «أ»، ومع ذلك، فكلا الرهانين سيكون لهُما منفعة مُتساوية.

في ذلك الوقت، كان تقديم دانييل برنولي لمفهُوم المنفعة؛ تلك الصِّفة الخفيَّة، لتفسير السُّلوك الإنساني عبر نظرية رياضية، هو طرح عجيب في بابه. ومما زاده روعةً حقيقة أنَّ قِيَم المنفعة للرهانات والجوائز المُتباينة لا تُلْحظُ مُباشرةً، على عكس القيم النَّقديَّة، بل

«تُستنتَج» عوضًا عن ذلك من «التَّفضيلات» التي يُبْديها المرء. وسيمضي على هذه الفكرة قرنان من الزَّمان قبل أن تُستوْعب دلالاتها استيعابًا كاملًا وتصير مقبولةً على نطاقٍ واسع بين علماء الإحصاء والاقتصاد.

في منتصف القرن العشرين، نشر جون فون نيومان (وهو عالمُ رياضياتٍ شهيرٌ سُمِّيت بنية أجهزة الكمبيوتر القياسية على اسمه)،[16] بالتَّعاون مع أوسكار مورجينسترن أساسًا «بديهيًّا» لنظرية المنفعة.[17] وما يعنيه ذلك الأساس هو كما يلي: طالما أنَّ التَّفضيلات التي يُبْديها فرد ما تُوفِّي قدرًا مُعيَّنًا من البديهيات الأساسيَّة الواجب على أي كيانٍ عقلاني أن يُوفِّيها، حينها «بالضرورة» يُمكن وصف اختيارات هذا الفرد بأنَّها تزيد للحد الأقصى القيمة المُتوقَّعة لدالَّة المنفعة. باختصارٍ: «أي كيانٍ عقلانيٍّ عليه أن يتصرَّف بُغية أن يزيد المنفعة المُتوقَّعة إلى أقصى حد».

ومهما طال الحديث عن أهمية هذا الاستنتاج فلن نُوفِّيه حقَّه. فبطُرقٍ شتَّى، كان مجال الذكاء الاصطناعي، وما يزال، مُتمحورًا على نحو أساسي حول اكتشاف أسرار وتفاصيل كيف نبني آلاتٍ عقلانية.

هيا بنا نُلقي نظرةً مُتعمِّقةً أكثر حول البديهيات التي يُتوقَّع من الكيانات العقلانية أن تُوفِّيها. إليك أوَّلُها؛ والتي تُسمَّى «التَّعدِّي». ومعناها أنَّك إذا كنت تُفضِّل «أ» على «ب»، وفي الوقت ذاته تُفضِّل «ب» على «ج»، إذن أنت تُفضِّل «أ» على «ج». يبدو هذا أمرًا بديهيًّا تمامًا! (إذا كنت تُفضِّل بيتزا السُّجُق على بيتزا الجُبن، وفي نفس الوقت تُفضِّل بيتزا الجُبن على بيتزا الأناناس، فمن المنطقي أن نُخمِّن أنَّك ستختار بيتزا السُّجُق وتترك بيتزا الأناناس.) وإليك ثانية هذه البديهيات والتي تُسمَّى «الرتابة». وهي أنَّك إذا كنت تُفضِّل الجائزة «أ» على الجائزة «ب»، وكنت مُخيَّرًا بين بطاقتَي يانصيب حيث «أ» و«ب» هما فقط النَّتيجتان المُحتملتان، فأنت ستُفضِّل البطاقة ذات الاحتمالية الأعلى لربح الجائزة «أ» عوضًا عن الجائزة «ب». ومرَّة أخرى، يبدُو هذا أمرًا غايةً في البداهة.

ولا تنحصر التَّفضيلات في أنواع البيتزا وبطاقات اليانصيب ذات الجوائز المالية فقط، بل تكون في سائر الأشياء مُطلقًا؛ وقد تكون متعلقةً بحيوات الآخرين والحياة المُستقبليَّة بالكامل على وجه الخُصُوص. وعند مُعالجة تفضيلاتٍ تنطوي على تتابُعٍ للأحداث مع مرور الوقت، غالبًا ما يُفرض افتراض إضافي يُسمَّى «الثَّبات»؛ ومعناه أنَّه إذا استهلَّ خطَّان مُستقبليان بنفس الحدث، وكان أحدهما «أ» والآخر «ب»، وكنت تُفضِّل «أ» على «ب»، فستظلُّ مُتمسِّكًا بتفضيلك لـ «أ» على «ب» حتى بعد انتهاء الحدث. قد يتراءى لك

أنَّ هذا الافتراض بديهي، لكنَّك قد تُفاجأ بما يترتَّب عليه من نتائج؛ فالمنفعة من أي سلسلةٍ من الأحداث هي مجموع المكافآت المرتبطة بكُلِّ حدثٍ من تلك الأحداث (والذي قد يتضاءل بمرور الوقت جراء نوعٍ من مُعدَّلات الاهتمام العقلي).[18] ومع أنَّ هذا الافتراض بأن «المنفعة هي مجموع المكافآت» ينتشر انتشارًا واسعًا – ويعُود في أصله على أقل تقديرٍ إلى نظرية «حساب اللذَّة» التي وُضعت في القرن الثَّامن عشر على يد مُؤسِّس مذهب النَّفعيَّة جيرمي بنثام؛ فإنَّ افتراض الثَّبات الذي يقوم عليه لا يُعدُّ صفةً لازمةً للكيانات العقلانية. فافتراض «الثَّبات» ينفي احتمالية أن تفضيلات المرء قد تتغيَّر بمرور الوقت، وهو ما يُخالف واقعنا المُشاهد.

ومع ما تحملُه تلك الأسس البديهيَّة من معقولية وما ترتَّب عليها من استنتاجاتٍ مُهمَّة، فإنَّ نظرية المنفعة قد لاقت ريحًا عاصفًا لا تهدأ من الاعتراضات مُنذُ أن بدأ صيتُها يذيع وتشتهر. فبعضُ الناس كان يزدريها لمظنَّة أنَّها تختزل الحياة في المال وحُبِّ الذَّات لا غير. (وقد وُصمت النظرية بأنَّها «أمريكية» استهزاءً وتهكُّمًا على لسان بعض الباحثين الفرنسيين،[19] رغم ما لها من جُذُورٍ في فرنسا.) في الحقيقة، تُعدُّ الرَّغبة في عيش حياةٍ فيها نُكران الذَّات والتَّخفيف من معاناة الآخرين هي غايتها الأسمى، أمرًا عقلانيًّا تمامًا. فالغيرية ما هي إلا أن يُقام لمصلحة الآخرين وسعادتهم وزن جوهري عند تقييم أي تفضيلاتٍ مستقبلية.

ثُمَّ هبَّت عاصفة أخرى من الاعتراضات حول صُعُوبة الحصول على الاحتمالات الضَّرُوريَّة وقيم المنفعة فضلًا عن ضربهما معًا لحساب المنافع المُتوقَّعة. أعتقد أن هذه الاعتراضات قد خلطت بين أمرَين؛ وهما: اختيار التَّصرُّف العاقل واختياره استنادًا إلى «حساب منافعه المُتوقَّعة». ومثال ذلك أنَّك إذا حاولت أن تفقأ إحدى مُقلتي عينيك بإصبعك، فإنَّك تجد جفنك قد انطبق ليحمي عينك؛ هذا تصرُّف عقلاني، ومع ذلك لم تتخلَّله أي حساباتٍ للمنفعة المُتوقَّعة. أو لنفترض جدلًا أنَّك تقُود دراجةً دون مكابح باتجاه سفح تلٍّ وأمامك خياران؛ إما أن تصطدم بجدارٍ إسمنتي وأنت بسُرعة عشرة أميالٍ في الساعة، وإما أن تصطدم بجدارٍ إسمنتي آخر على مسافةٍ أبعد وأنت بسُرعة عشرين ميلًا في الساعة، فأيُّ الجدارَين ستختار؟ إذا كان اختيارك هو أن تصطدم وأنت بسُرعة عشرة أميالٍ في الساعة، فإليك تهنئتي! هل تخلَّل قرارك أيُّ حساباتٍ للمنافع المُتوقَّعة؟ على الأرجح لا، ومع ذلك فإنَّ اختيار الاصطدام بسُرعة عشرة أميالٍ في الساعة لا يزال يُوصفُ بالاختيار العقلاني. وهذا نابع من افتراضَين أساسيَّين؛ أوَّلُهُما أنَّك آثرتَ

الجراح الأخف على الجراح الأشد، وثانيهما أنَّ تزايد سُرعة الاصطدام يزيد من احتماليَّة أن تتخطَّى مستوى أي جُرُوحٍ مُتوقَّعة مهما زاد سُوءُه. ومن هذَين الافتراضَين نخلُص إلى أنَّه رياضيًّا، ودون التَّطرُّق إلى أيِّ أرقامٍ مُطلقًا، الاصطدام بسُرعة عشرة أميالٍ في الساعة له منفعة مُتوقَّعة أعلى من الاصطدام بسُرعة عشرين ميلًا في الساعة.[20] وخُلاصة القول هي أنَّ تعظيم المنفعة المُتوقَّعة إلى أقصى حدٍّ قد لا يتطلَّبُ حساباتٍ لأيِّ توقُّعاتٍ أو منافع؛ فهذا الأمر لا يعدُو كونه محض توصيفٍ ظاهريٍّ للكيانات العقلانية.

نقد آخر لنظرية العقلانية يكمُنُ في تحديد محلِّ اتِّخاذ القرارات. بصيغةٍ أخرى، ما الأشياء التي تُعدُّ كيانًا؟ أظنُّ أنَّنا نتَّفق على أنَّ البشر كيانات، ولكن ماذا عن الأُسَر والقبائل والشَّركات والثَّقافات والأمم القوميَّة؟ إذا ما تأمَّلنا بعض الحشرات الاجتماعية كالنَّمل مثلًا، فهل يُعقل أن نعتبر أي نملة بمفردها كيانًا ذكيًّا، أم أنَّ الذكاء يكمُن حقًّا في المُستعمرة بأسرها كوحدةٍ واحدةٍ تتكوَّن من دماغٍ ضخمةٍ مُؤلَّفةٍ من العديد من أدمغة وأجساد النَّمل التي يربطُها معًا نظامُ تواصُلٍ بإفراز الرَّوائح (الفرمُونات) بدلًا عن نظامٍ يعتمد على الإشارات الكهربيَّة؟ من وجهة نظرٍ تطوُّريَّة، هذا التَّصوُّر حول النَّمل رُبَّما يكون أجدى من غيره؛ لما كان بين النَّمل عادةً من ترابُطٍ وثيقٍ في أي مُستعمرةٍ. يبدو أن النَّمل وغيره من الحشرات الاجتماعية يفتقر، كأفرادٍ، إلى غريزةٍ للحفاظ على الذَّات باعتبارها غريزةً مُنفصلةً عن غريزة الحفاظ على المُستعمرة. فهُو دائمًا ما يهُبُّ لخوض المعارك ضدَّ الغُزاة، حتى ولو كان موتُهُ مُحتَّمًا. بيد أنَّنا نرى أحيانًا بعض البشر يفعلون الشَّيء ذاته ليُدافعوا عن غيرهم من البشر وإن كانوا غير أُولي قُربى؛ كأنَّ النوع بأكمله يستفيد من وجود عددٍ ضئيلٍ من أفراده لديهم الاستعداد للتَّضحية بأنفسهم في المعارك أو الذهاب في رحلاتٍ بحريَّةٍ استكشافيَّةٍ جامحةٍ تحفُّها المخاطر من كُلِّ جانب، أو تنشئة وتربية نسل أناسٍ آخرين. في هذه الحالات، إذا نظرْنا إليها بعينٍ تُحلِّل نظرية العقلانية على أساسٍ فردي محضٍ، فإنَّنا لا محالة فاقدون عُنصُرًا جوهريًّا من الصُّورة الكاملة.

أما بقيَّة الاعتراضات الرَّئيسيَّة على نظرية المنفعة فهي اعتراضات تجريبية؛ أي إنها مبنيَّة على أدلةٍ تجريبيَّة تُشير إلى أنَّ الإنسان كائن لا عقلاني أصلًا. نحن نُخفق في الالتزام بالأسُس البديهيَّة بأساليب منهجية.[21] وغايتي هُنا ليست أن أُدافع عن نظرية المنفعة بوصفها نمُوذجًا رسميًّا للسُّلُوك البشري. في الواقع، لا يُمكن للبشر أن يتصرَّفُوا بعقلانيَّة؛ فتفضيلاتُنا تمتدُّ لتُؤثِّر في حيواتنا المُستقبليَّة بأكملها، بل وحيوات أبنائنا وأبناء

أبنائنا، وحيوات الآخرين الذين يعيشُون الآن أو سيعيشُون في المُستقبل. مع ذلك، فنحن نُخفقُ حتى في تحريك القطع على رُقعة الشطرنج على نحوٍ صحيح؛ تلك الرُّقعة التي تُمثِّل عالمًا صغيرًا وبسيطًا ذا قواعد مُحدَّدةٍ ومدًى غايةً في القصر. وهذا بالطَّبع ليس لأنَّ «تفضيلاتنا» لا عقلانيَّة، بل بسبب «تعقد» مُعضلة اتخاذ القرارات. فمقدار كبير من بِنيتنا المعرفيَّة موجُود لسدِّ الثَّغرة بين أدمغتنا الصَّغيرة والبطيئة وبين التَّعقيد الهائل على نحوٍ غير مفهوم لُمعضلة اتخاذ القرارات التي نُواجهُها في كُلِّ حين.

وهكذا، رغم أنه من غير المعقول أن نبني نظريةً عن الذكاء الاصطناعي النافع استنادًا إلى افتراض أنَّ البشر كيانات عقلانيَّة، فسيكون من الصَّواب أن نفترض أنَّ الإنسان البالِغ الراشد غالبًا ما يكون لدَيه تفضيلات مُتَّسقة بخُصُوص حياته المُستقبليَّة. وبيان ذلك هو أنَّك «إذا قُدِّر لك بطريقةٍ ما واستطعت أن تُشاهد فيلمَين يصف كُلُّ واحدٍ منهما مسيرة حياةٍ مُستقبليَّةٍ بإمكانك أن تعيشها لو أردت وصفًا دقيقًا مُتأنِّيًا يجعلُك تعيشُ أجواءها كأنَّها حقيقة، تستطيع أن تختار أيهما تُفضِّل أو تُعبِّر عن أن كليهما إليك سواء».[22]

لعلَّ هذا الادِّعاء أقوى مما نحتاج إذا كانت غايتُنا الوحيدة هي أن نضمن أنَّ الآلات ذات الذَّكاء الكافي لن تكُون جالبةً للنَّكبات على الجنس البشري. ومفهُومُ النَّكبة هذا يستلزم حياةً غير مفضلة بلا شك. ولنتفادى النَّكبات، علينا فقط أنْ نحصُر ادِّعاءنا هذا على أنَّ الإنسان البالغ الراشد يقدر على تمييز المُستقبل المنكُوب حين يُطرح أمامه بتفصيلٍ كبير. وبلا شك، فإنَّ التفضيلات البشرية لها بنية أكثر دقَّة، وربما أكثر قابلية للتَّثبُّت منها، من مجرد مُفاضلة بين أبيض أو أسود؛ «عالم بلا نكباتٍ أفضل من عالمٍ منكُوب».

في الحقيقة، يُمكن لنظرية للذكاء الاصطناعي النافع أن تتَّسع لتحتوي عدم الاتساق في تفضيلات البشر، لكنَّ ذلك الجزء غير المُتسق في تفضيلاتك لن يجري إرضاؤه أبدًا، وحينها لن يكون في جعبة الذكاء الاصطناعي شيءٌ ليُقدِّمه. دعنا نفترض على سبيل المثال أنَّ تفضيلاتك للبيتزا تُخالف أحد الأسُس البديهية؛ وهو «التَّعدِّي»:

الروبوت: عود حميد! أتريد بعضًا من بيتزا الأناناس؟
أنت: كلا! كان عليك أن تعرف أنِّي أُفضِّل بيتزا الجُبن على بيتزا الأناناس.
الروبوت: حسن، سأجهِّز لك بيتزا جُبن حالًا!
أنت: لا، شكرًا لك. أنا أحبُّ بيتزا السُّجُق أكثر.
الروبوت: معذرة، سأجهِّز لك بيتزا سجُق.
أنت: في الحقيقة أنا أُفضِّل بيتزا الأناناس على بيتزا السُّجُق.

الروبوت: هذا خطئي، لتكن بيتزا الأناناس إذن!
أنت: لقد قُلتُ لتوِّي إني أُفضِّل بيتزا الجُبن على بيتزا الأناناس.

على هذا المنوال، مهما جهَّز الروبوت من أنواع البيتزا فلن يُرضيك أو يُلبِّي رغبتك؛ لأنَّ هناك دائمًا بيتزا أخرى تُفضِّلُها على ما سيُقدَّم إليك. لكنَّ الروبوت قادر على تلبية الجزء المُتَّسق من تفضيلاتك فقط؛ لنفترض مثلًا أنَّك تُفضِّل أن تأكل أي نوعٍ من أنواع البيتزا الثلاثة على ألا تأكل بيتزا إطلاقًا. في هذه الحالة، الروبوت النافع سيُجهِّز لك أي نوعٍ من الأنواع الثلاثة التي تفضلها من البيتزا، وحينها سيكون قد لبَّى رغبتك في عدم ترك أكل البيتزا، ثمَّ يترُكُك لتتفكَّر بترويَّةٍ في تفضيلاتك غير المتسقة على نحوٍ مزعج لنوعيَّة الإضافات على البيتزا.

(١-٤) عقلانيَّة الجماعة

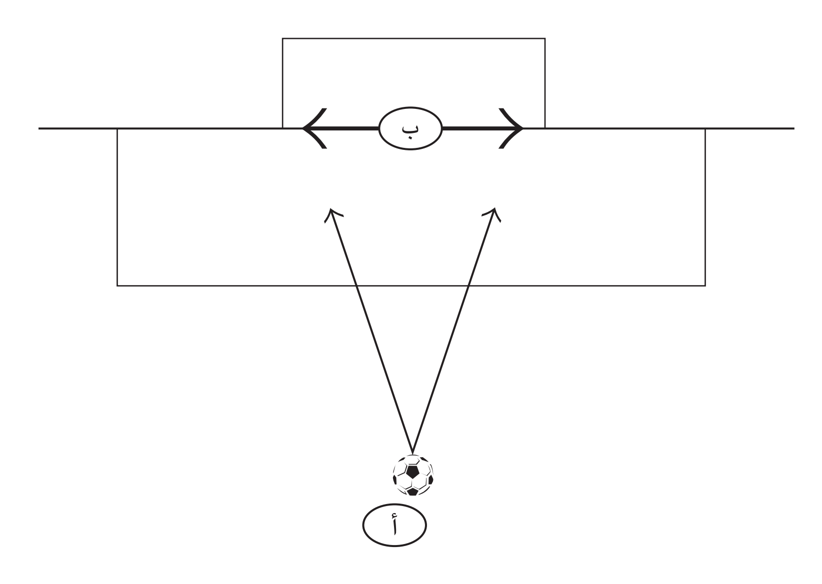
الفكرة الأساسية التي تقضي بأنَّ الكيان العقلاني يتصرَّف ليزيد من المنفعة المُتوقَّعة إلى أقصى حدٍّ، هي فكرة بسيطة بالقدر الكافي، حتى ولو أنَّ تنفيذها فعليًّا يُعدُّ أمرًا بالغ التعقيد حتى يكاد يكون مستحيلًا. لكن هذه النظرية تصلُحُ فقط لتفسير الحالات التي يكون فيها كيانٌ واحد يتصرَّف بمُفرده. أما إنْ كانوا أكثر من كيانٍ، فإنَّ ذلك التصور، الذي يرى أنه يُمكنُنا ولو نظريًّا تحديد احتمالات النتائج المختلفة لتصرُّفات الفرد، يُصبح إشكاليةً مُعقدة. والسَّبب وراء ذلك هو أنَّ هناك جزءًا ما من العالَم، وهو الكيان الآخر، يُحاول الآن أن يُخمِّن كُنْه التصرفات التي ستقوم بها، والعكس صحيح، وهكذا، فلا نرى سبيلًا واضحًا لتحديد احتمالات ما سيصْدُر عن ذلك الجزء من العالَم من تصرُّفات. وبدون الاحتمالات فإنَّ تعريف التَّصرُّف أو الفعل العقلاني بأنَّه يهدف إلى زيادة المنفعة المُتوقَّعة إلى أقصى حدٍّ، يكون غير قابلٍ للتطبيق.

وحالما ينضمُّ شخص آخر إلى العمليَّة، فإنَّ على الكيان أن يجد طريقةً أخرى لاتِّخاذ القرارات العقلانية. وهنا يأتي دور «نظرية الألعاب». لا يغُرنَّك الاسم؛ فهي ليست بالضرورة تتمحور حول الألعاب بالمعنى التقليدي، بل هي تصوُّر عام يُحاول بسْط فكرة العقلانية إلى الحالات التي تضُمُّ أكثر من كيانٍ واحد. وهذا مُهم على نحو واضح لتحقيق غاياتنا؛ لأننا لا نُخطِّط (حتى الآن) لبناء روبوتات لنُرسلها للعيش على كواكب غير مأهولةٍ

في نُظُم نجميَّة بعيدة؛ بل على العكس تمامًا، نحن نبني روبوتات لنستخدمها في عالمنا الذي نسكنه نحن البشر.

ولإيضاح فائدة نظرية الألعاب وحاجتنا إليها، إليكم المثال البسيط التالي: أليس وبوب يلعبان كرة القدم في حديقة منزلهما الخلفية (انظر الشَّكل ٢-١). أليس تستعد للعب ضربة جزاءٍ وبوب يقف حارسًا للمرمى. وهي بين خيارَين؛ إما أن تُسدِّد الكرة على يمين بوب أو شماله. ولأنَّها يمينية القدم، فمن الأسهل لها إلى حدٍّ ما والأدقِّ أيضًا أن تُسدِّد الكرة إلى يمين بوب. ولأنَّ أليس ركلتُها سريعة وخاطفة، يعرف بوب أنَّ عليه أن يختار أن يندفع إما يمينًا أو شمالًا على الفور؛ فهو لن يحظى بالوقت الكافي لينتظر ويرى في أيِّ اتجاهٍ ستذهب الكرة. وقد يُفكِّر بوب على هذا النَّحو: «أليس لدَيها فرصة طيبة لتسجيل الهدف إن سدَّدت الكرة إلى يميني لأنَّها يمينية القدم، لذلك أظنُّ أنها ستختار هذا وسأندفع أنا يمينًا.» لكنَّ أليس ليست بالساذجة وتقدر على تصوُّر طريقة تفكير بوب تلك، ولذلك ستختار أن تُسدِّد إلى شمال بوب. لكنَّ بوب ليس بالساذج ويقدر على تصوُّر طريقة تفكير أليس تلك، ولذلك سيندفع شمالًا. لكنَّ أليس ليست بالساذجة وتقدر على تصوُّر طريقة تفكير بوب تلك ... وهكذا دواليك، أظنُّ أنَّ الأمر قد اتَّضح. ولنُلخِّص الأمر بطريقةٍ أخرى، إذا كان هناك خيار عقلاني أمام أليس لتتَّخذه، فبإمكان بوب أن يتصوَّره هو الآخر وأن يتوقَّع حدوثه ويمنعها من تسجيل الهدف، لذلك فالاختيار لا يمكن أن يكون عقلانيًّا منذ البداية.

في وقتٍ مُبكِّرٍ من التاريخ، وتحديدًا بحُلُول عام ١٧١٣، اكتُشف حلٌّ لهذا اللغز، مرة أخرى من خلال تحليل ألعاب المقامرة.[23] الحيلة هنا ليست أن تختار تصرُّفًا مُعينًا، ولكن أن تختار «خُطَّةً عشوائية». ومثال ذلك هو أنَّ أليس يمكنها أن تختار الخُطَّة التالية: «التسديد إلى يمين بوب باحتمالية تسجيل بنسبة ٥٥ بالمائة، أو التسديد إلى شمال بوب باحتمالية تسجيل بنسبة ٤٥ بالمائة». أما بوب فيُمكنه أيضًا انتهاج الخُطَّة التالية: «الاندفاع إلى اليمين باحتمالية صدٍّ بنسبة ٦٠ بالمائة، أو إلى الشِّمال باحتمالية صدٍّ بنسبة ٤٠ بالمائة». كلاهُما يرمي في ذهنه عُملةً معدنيةً متحيزة على نحوٍ ملائم مباشرة قبل أن يتصرَّفا لكيلا يُبديا نواياهما. بالتَّصرُّف «على نحو غير مُتوقَّع»، يتجنَّب كُلٌّ من أليس وبوب التَّضارُبات التي شهدناها في الفقرة السابقة. وحتى إن علم بوب بخُطَّة أليس العشوائية بطريقةٍ ما، فلن يُفيده هذا بشيءٍ إلا إذا كان يملك بلورة العرَّافين السِّحريَّة.



شكل ٢-١: أليس تستعد للعب ضربة جزاءٍ على مرمى بوب.

والسؤال التالي الذي يطرحُ نفسه هو: ما هي الاحتمالات؟ وهل خُطَّة أليس التي اختارتها وفيها نسبة ٥٥ بالمائة مقابل نسبة ٤٥ بالمائة، تُعتبر خُطَّةً عقلانية؟ في الحقيقة، تتوقَّف القِيَم الدقيقة على مدى دقَّة أليس وهي تُسدِّد الكرة إلى يمين بوب، كما تتوقَّف على مدى براعة بوب في التصدي للكرة وهو يندفع إلى الاتجاه الصحيح، وغير ذلك. (طالع قسم «الملاحظات» لتقف على التحليل الكامل.)[24] ومع ذلك، فالمعيار العام غاية في البساطة:

(١) أن تكون خُطَّة أليس هي أفضل ما جادت به قريحتُها، بافتراض أنَّ خُطَّة بوب ثابتة.

(٢) أن تكون خُطَّة بوب هي أفضل ما جادت به قريحتُه، بافتراض أنَّ خُطَّة أليس ثابتة.

إذا تحقَّق ذلكُما الشَّرطان، حينها نقول إن كلتا الخُطَّتَين في حالة توازن. ويُسمَّى هذا النوع من التوازن بـ «توازن ناش»، تخليدًا لذكرى العالم جون ناش الذي استطاع عام ١٩٥٠ وهو بسنِّ الثانية والعشرين أن يُثبت وجود هذا التوازن بين أيِّ عددٍ من الكيانات مع وجود أي تفضيلات عقلانية ومهما كانت قوانين اللعبة. وبعد أن صارع

جون ناش مرض انفصام الشخصية لعدة عقودٍ، تغلَّب عليه أخيرًا وتعافى، ومُنح جائزة نوبل التذكارية في الاقتصاد عام ١٩٩٤ نظير اكتشافه ذلك.

بالنسبة لمُباراة كرة القدم بين أليس وبوب، فإننا نجد توازنًا واحدًا فقط. في حالاتٍ أخرى، ربُما توجد عدَّة توازنات، ولذلك فإن مفهوم توازنات ناش، على عكس ذلك الخاص بقرارات المنفعة المُتوقَّعة، لا تُرشدُنا دومًا إلى الطريق الأمثل للتَّصرُّف.

والأسوأ من ذلك، أنَّ هناك مواقف نجد فيها أن توازن ناش يبدو أنه يقُودُنا إلى نتائج غير مرغوبٍ بها على نحوٍ كبير. ومن أمثلة هذه المواقف ما اشتُهر باسم «مُعضلة السُّجناء»، والتي سماها بهذا الاسم ألبرت تاكر عام ١٩٥٠؛ وهو المُشرفُ على أُطرُوحة جون ناش لرسالة الدكتوراه.[25] دعونا نُوضِّح أنَّ اللعبة هي نمُوذج مجرد لتلك المواقف الشائعة جدًّا في الحياة الواقعية حين يكون التعاون المشترك هو أفضل خيارٍ لكل الأطراف المعْنيَّة، لكن على الرغم من ذلك يختارون أن يُدمِّر بعضُهم بعضًا.

وبيان مُعضلة السُّجناء هذه كما يلي: أليس وبوب مُشْتبه بهما في جريمةٍ ما ويُحقَّقُ معهما على حدة. وكلاهما أمامه اختيار؛ إما أن يعترفا للشرطة ويشي كُلُّ واحدٍ بشريكه في الجريمة، وإما أن يلزما الصمت.[26] فإن لزم الاثنان الصمت، ستُوجَّه إليهما تُهم هيِّنةً ويقضيان سنتَين في السجن، وإن اعترف كلاهما ووشى كُلُّ واحدٍ بصاحبه، سيُدانان بتُهمٍ خطيرةٍ ويقضيان عشْر سنين في السجن. أما إذا اعترف أحدهما ولزم الآخر الصمت، فسيُطلق سراحُ من اعترف ويُسجن شريكه مُدَّة عشرين سنة.

في تلك الحالة، ستُفكِّر أليس كما يلي: «إن كان بوب سيعترف أمام الشُّرطة، فعليَّ أن أعترف أنا أيضًا (فعشْر سنواتٍ أهون من عشرين)؛ أما إن كان سيلزم الصَّمت، فلأعترفنَّ أنا (فالحُرِّية أفضل من قضاء سنتين في السجن)؛ إذن في كلتا الحالتين، عليَّ أن أعترف.» وكذلك سيُفكِّر بوب بنفس الطريقة. لذا ينتهي المطافُ وقد اعترف كلاهما بالجُرم وعُوقبا بالسَّجن عشْر سنين، رُغم أنَّهُما كانا سيقضيان سنتين فقط إذا لزما الصمت معًا. والمشكلة هنا أنَّ التزام الصمت المشترك لا يُحقِّق توازن ناش؛ لأنَّ كلَّ واحدٍ منهما لديه من الباعث ما يدفعه لينقلب على صاحبه ويعترف ليفوز بالحرية.

لاحظ أنَّ أليس كان بإمكانها أن تُفكِّر كما يلي: «أيَّما طريقة أفكر بها، فسيُفكِّر بها بوب أيضًا، هكذا سينتهي بنا المطافُ وقد اخترنا القرار ذاته. وطالما أنَّ الصمت المشترك أفضل من اعتراف أحدنا على الآخر، فعلينا إذن أنْ نرفض الاعتراف وأنْ نلزم الصمت.» يُسلِّم نمط التَّفكير هذا بأنَّ كُلًّا من أليس وبوب، بوصفهما كيانين عقلانيَّين، سيتَّخذان

قراراتٍ تصُبُّ في مصلحتهما المُشتركة لا قراراتٍ فردية بحْتة. هذا منهج من مناهج كثيرةٍ حاول علماء نظرية الألعاب أن يتَّبعوها لعلَّهم يصلُون إلى حلول أقل إحباطًا لمُعضلة السُّجناء هذه.[27]

ومثال آخر شهير على توازنٍ يُحقِّق نتائج غير مرغوب فيها هو «مأساة المشاع» التي حُلِّلت تفاصيلها للمرة الأولى عام ١٨٣٣ على يد الاقتصادي الإنجليزي ويليام لويد،[28] لكنَّ عالم البيئة جاريت هاردن هو من سمَّاها وقدَّمها عام ١٩٦٨ حيث نالت اهتمامًا عالميًّا.[29] وهذه المأساة تظهر عندما يتشارك جمع من الناس في استهلاك موردٍ مُشترك يتجدَّد ببُطءٍ كأراضي الرَّعي أو مخزون سمكي في حيِّز مائي. وفي غياب الرادع الاجتماعي أو القانوني، فإن التَّصرف الوحيد الذي يُحقِّق توازن ناش بين الكيانات الأنانية (التي لا تهتم بمصلحة غيرها)، هو أن يستهلكوا ذاك المورد قدر المُستطاع مما يتسبَّب في نفاده سريعًا. والحلُّ الأمثل، والمُتمثل أن يتشارك الجميع استهلاك المورد ليكون إجمالي استهلاكهم مُستدامًا، لا يُحقِّق توازنًا لأنَّ كل فردٍ لديه ما يدفعه للغشِّ واستهلاك أكثر من الحصَّة العادلة ليتحمَّل الآخرون كُلفة جشعه. عمليًّا، بالطبع، البشر قادرون أحيانًا على تفادي حدوث هذه المأساة بوضع آليات مثل تحديد الحصص وفرض العقوبات ووضع نُظُم التَّسعير. وقُدرتُهم على فعل ذلك تنبُع من كونها غير مقصورةٍ على تقرير حصَّة الاستهلاك، بل بإمكانهم أيضًا أن يُقرِّروا «التَّواصل» بعضهم مع بعضٍ. وبتوسيع مشكلة اتخاذ القرار على ذلك النَّحو، فإننا نجد حلولًا تُناسب الجميع وتصبُّ في مصلحتهم.

تلك الأمثلة وغيرها الكثير، إنما تُوضِّح حقيقة أنَّ توسيع نطاق نظرية القرارات العقلانية لتشمل كيانات متعدِّدة يُنتج عددًا مهُولًا من السُّلوكيات المُعقَّدة والمُثيرة للانتباه. كما أن هذا ذو أهمية شديدة في الوقت ذاته؛ لأنه كما أظنُّ أنَّه شديد الوضوح، أنَّ هناك أكثر من إنسانٍ في العملية. وعما قريبٍ ستُشاركنا الآلات الذكية هي الأخرى فيها. ولا حاجة بي أن أُنبِّه إلى ضرورة السَّعي إلى تحقيق تعاونٍ مُشتركٍ تكون ثمرتُه هي مصلحة البشر، عوضًا عن اختيار أن يُفني أحدنا الآخر.

(٢) أجهزة الكمبيوتر

المُكوِّن الأول لإنشاء آلاتٍ ذكيةٍ هو أن يكون لدينا تعريف صائب لماهيَّة الذكاء. أما المُكوِّن الثاني فهو الآلة التي يُمكن أن تُحقِّق هذا التَّعريف. ولأسبابٍ سرعان ما ستتَّضِح فيما

بعد، فالآلة هنا هي جهاز الكمبيوتر. كان يُمكن لها أن تكون شيئًا آخر — فعلي سبيل المثال، كان يمكن لنا أن نُحاول بناء آلاتٍ ذكيةٍ عن طريق بعض التفاعلات الكيميائية المعقَّدة أو السيطرة على الخلايا الحية[30] — ومع ذلك، فإن الأجهزة التي صُمِّمت لعمليات الحوسبة، بداية من الآلات الحاسبة الميكانيكية المبكرة جدًّا فصاعدًا، لطالما بدت لمُخترعيها على أنَّها المُستقَر المُناسب للذكاء.

إننا، في وقتنا الحالي، اعتدنا أجهزة الكمبيوتر في حياتنا، حتى إننا بالكاد نلتفت إلى قدراتها الخارقة. إن كنت تمتلك جهاز كمبيوتر محمولًا أو مكتبيًّا أو هاتفًا ذكيًّا، فتمعَّن في أيٍّ منها؛ ستجده صندوقًا صغيرًا ذا وسيلةٍ ما لكتابة الرُّمُوز. بالرموز التي تُدخلها فقط، يُمكنُك أن تُنشئ برامج تجعل من هذا الصندوق شيئًا جديدًا؛ رُبَّما شيئًا سحريًّا ينسج مشهدًا مُكونًا من صورٍ متحركة لبواخر عابرةٍ للمُحيطات وهي تصطدم بجبالٍ جليدية، أو لكواكب فضائيين طوال القامة زُرق البشرة؛ أدخل رُموزًا أكثر، وها هو ذاك الصندوق يُترجم من اللغة الإنجليزية إلى اللغة الصينية؛ أدخل رُمُوزًا أكثر، ويصير صندوقًا يسمعُك ويُحدِّثُك؛ أدخل رُمُوزًا أكثر، ليغلب بطل العالَم في لُعبة الشطرنج.

تلك القُدرة التي تُمكِّن صُندُوقًا واحدًا من تنفيذ أيِّ عمليةٍ يُمكنك تخيُّلها تُسمَّى «العموميَّة»، وهو مفهوم قدَّمه آلان تورينج لأول مرةٍ عام ١٩٣٦.[31] والعموميَّة تعني أنَّنا لسنا بحاجةٍ إلى آلةٍ مُستقلَّةٍ للحساب، وأخرى للترجمة الآلية، وثالثة للعب الشطرنج ورابعة لاستيعاب الكلام المنطوق، وخامسة لإنشاء الرُّسُوم المتحركة؛ لا! بل هي آلة واحدة تقدر على تنفيذ كل ما سبق. إن جهاز الكمبيوتر المحمُول خاصتك يُطابق في أسس عمله أي كمبيوتر في مصاف أجهزة الخوادم الضَّخمة التي تُديرها كبرى شركات تكنولوجيا المعلومات في العالم، وحتى تلك المُجهَّزة بوحدات مُعالجة التنسور ذات الإمكانيات العالية والمُخصَّصة لأغراض تعلُّم الآلة. كما أنَّه يُطابق في أسس عمله أي أجهزةٍ حاسُوبيةٍ ستُخترع مُستقبلًا. وبفرض أنَّ جهازك مُزوَّد بذاكرةٍ كافية، فإنَّه يقدر على تنفيذ نفس المهام بالضَّبط؛ لكنَّ الفارق أنَّه سيستغرق زمنًا أطول.

تُعدُّ الورقة البحثيَّة التي قدَّم فيها آلان تورينج مفهوم العمومية من أهم ما كُتب على الإطلاق. في ورقته تلك، كتب وصفًا لجهازٍ حاسوبي بسيطٍ يقدر على قبول توصيف أي جهاز حاسوبي آخر كمُدخلاتٍ، ثمَّ يعمل جنبًا إلى جنبٍ مع مُدخلات ذاك الجهاز الآخر ليُقدِّم نفس المُخرجات التي كان ليُخرجها، عن طريق مُحاكاة عمله من خلال مُدخلاته. نحن الآن نُسمِّي هذا الجهاز الأول «آلة تورينج العمومية». ولإثبات عموميتها، طرح

تورينج تعريفَين دقيقَين لنوعَين جديدَين من العناصر الرياضيَّة؛ وهما: الآلات والبرامج. يعمل هذان العنصران معًا لتعريف سلسلة من الأحداث؛ على وجه الخُصُوص، سلسلة من تغيُّرات الحالة في الآلة وذاكرتها.

إن اكتشاف عناصر رياضيةٍ جديدةٍ هو شيء نادر الحُدُوث في تاريخ الرِّياضيات. ففي فجر التاريخ المُدوَّن، بدأت الرِّياضيَّات بظهور الأعداد، ثمَّ حوالي سنة ٢٠٠٠ قبل الميلاد، اكتشف قدماء المصريين والبابليون العناصر الهندسية (النقاط، والخطوط، والزَّوايا والمساحات وهلُمَّ جرًّا) وعملوا بها. وفي سنوات الألفيَّة الأولى قبل الميلاد، قدَّم علماءُ الرِّياضيات الصِّينيون المصفُوفات، بينما المجمُوعات كعناصر رياضيَّة عُرفت مُؤخَّرًا في القرن التاسع عشر. ويُعدُّ العنصران الجديدان اللذان قدَّمهُما تورينج؛ الآلات والبرامج، أعظم العناصر الرِّياضيَّة التي اختُرعت على مرِّ العُصُور. ومن عجيب التَّقادير أنَّ علم الرِّياضيات قد أخفق إخفاقًا ذريعًا في إدراك عظمة هذَين العنصرَين الرِّياضيَّين، وابتداءً من أربعينيات القرن الماضي فصاعدًا، أُلحقت دراسة أجهزة الكمبيوتر والحوسبة بأقسام الهندسة في مُعظم الجامعات الرائدة.

ازدهر العلمُ الذي ظهر، وهو علم الكمبيوتر، خلال السَّبعين سنةً اللاحقة، وقدَّم مجموعةً كبيرةً وجديدةً من المفاهيم والتَّصاميم والأساليب والتَّطبيقات، كما تمخَّض عنه سبع من أهمِّ ثماني شركاتٍ في العالم.

المفهوم الرَّئيسيُّ في علم الكمبيوتر يكمُن في «الخوارزميَّة»؛ وهي تُعرَّف بأنَّها طريقة مُحدَّدة بدقَّةٍ شديدةٍ لحوسبة شيءٍ ما. وفي عصرنا هذا، نرى تلك الخوارزميات حولنا كأجزاءٍ مألوفةٍ من حياتنا اليومية؛ فمثلًا خوارزميَّة الجذر التربيعي في حاسبة جيبٍ آليةٍ تستقبل العدد كأحد المُدخلات ثمَّ تحسب الجذر التربيعي لذلك العدد وتُظهره كأحد المُخرجات؛ خوارزميَّةُ لعب الشطرنج تحلُّ محلَّ أحد اللاعبين وتنتظر لوضعها في اللعب ثم تُبادر بتحريك إحدى القطع؛ خوارزميَّةُ تحديد الطُّرُق تضع في حُسبانها موقع البداية وموقع الوصول وخريطة الطُّرُق ثمَّ تُخبرك بأسرع طريقٍ يصل بين نقطة البداية ونقطة الوصول. يُمكنُنا وصف الخوارزميات باستخدام اللغة كالإنجليزية أو باستخدام طُرُق التَّدوين الرياضي، ولكن إذا أردنا أن نُطبِّق خوارزميَّةً ما فعلينا كتابتُها كبرامج باستخدام إحدى «لُغات البرمجة». وتُصمَّم الخوارزميات الأكثر تعقيدًا باستخدام خوارزميات أبسط كوحدات بنائيَّةٍ تُسمَّى «الروتينات الفرعيَّة». ومثال ذلك هو السَّيارة الذاتية القيادة

التي قد تستخدِم خوارزميَّة تحديد الطُّرُق كروتينٍ فرعي لمعرفة اتِّجاهات سيرها. وبهذه الطَّريقة تُبنى النظم البرمجيَّة البالغة التَّعقيد، طبقةً تلو الأخرى.

ومسألة المكوِّنات المادية لأجهزة الكمبيوتر تُهمُّنا أيضًا؛ لأنَّ أجهزة الكمبيوتر الأسرع ذات الذاكرة الأكبر تُتيح للخوارزميات أن تُشغِّل أسرع وأن تُعالج معلوماتٍ أكثر. والتَّقدم في هذا المجال معروف لكنَّه مُدهش. إن أول جهاز كمبيوتر إلكتروني قابل للبرمجة طُرح للبيع التجاري، «فيرانتي مارك ١»، كان يُمكنه تنفيذ نحو ألف ($١٠^{٣}$) أمرٍ في الثانية الواحدة وكان مُزوَّدًا بما يقرُب من ألف بايت من الذَّاكرة الرئيسية. أما أسرع جهاز كمبيوتر في أوائل ٢٠١٩، وهو «ساميت» بمُختبر أوك ريدج الوطني في ولاية تينيسي، فهو يُعالج نحو $١٠^{١٨}$ أمرًا في الثانية الواحدة (أي أسرع بألف تريليون مرة)، ومُزوَّد بذاكرةٍ سعتها $٢٫٥ \times ١٠^{١٧}$ بايت (أي أكبر بـ ٢٥٠ تريليون مرة). وهذا التَّقدم إنما هو ثمرةُ الجهود المبذولة في مجال الأجهزة الإلكترونية وحتى في الأمور الفيزيائية الكامنة وراءها والتي فتحت أبوابًا شتَّى أمام تقنية التَّصغير التصميمي.

ورُغم أنَّ المُقارنات بين الكمبيوتر والعقل البشري ليست ذات معنًى في هذا المقام، لكنَّ قُدرات الكمبيوتر «ساميت» قد فاقت قليلًا قدرات العقل البشري والتي كما ذكرنا آنفًا، تُقدَّر بما يقرُب من $١٠^{١٥}$ مشابك عصبية، و«زمن دورة» يصل إلى جزء من مائة من الثانية، مقارنة بحدٍّ أقصى نظري يصل لقُرابة $١٠^{١٧}$ «عملية» في الثانية الواحدة. والفارق الجوهري بين الاثنين يكمن في الطاقة المُستهلكة؛ فكمبيوتر «ساميت» يستهلك طاقةً أكثر بمليون مرة من العقل البشري.

«قانون مور»، والذي هو إحدى الملاحظات التجريبية التي تقول إن عدد المُكوِّنات الإلكترونية الموجودة في الرقاقات يتضاعف كل سنتين، يُتوقَّع أنَّ يظلَّ ساريًا حتى عام ٢٠٢٥ أو نحو ذلك، ولكن بمُعدَّلٍ أبطأ قليلًا. لسنواتٍ عديدة، أعاقت الحرارة العالية الناتجة عن التَّبديل السريع لترانزستورات السيليكون السُّرعات العالية لأجهزة الكمبيوتر، وعلاوة على هذا، لا يُمكنُنا تصغير حجم الدوائر الكهربية أكثر مما هي عليه الآن؛ فالأسلاك والمُوصِّلات، طبقًا لعام ٢٠١٩، لا يتعدَّى عرضها أكثر من خمسٍ وعشرين ذرَّة، ويتراوح سُمكُها بين خمس وعشر ذرات. وفي ما بعد عام ٢٠٢٥، سنحتاج إلى استخدام ظواهر فيزيائيةٍ أكثر تطوُّرًا؛ بما في ذلك أجهزة المُواسعة السالبة،[32] والترانزستورات الأحاديَّة الذرة، وأنابيب الجرافين النانوية، والضَّوئيات؛ وذلك للحفاظ على وتيرة التَّطور التي يتنبأ بها قانون مور (أو أي قانونٍ آخر يخلُفُه).

وثمَّة سبيل آخر بدلًا من زيادة سُرعة أجهزة الكمبيوتر المُتعدِّدة الاستعمالات، والذي يتمثَّل في أن نبني أجهزة ذات غرضٍ مُحدَّدٍ مُعدَّة لتُعالج نوعًا واحدًا من عمليات الحوسبة. على سبيل المثال، وحدات معالجة التنسور التي صمَّمتها جوجل تهدف للقيام بالعمليات الحسابية المطلوبة لخوارزميات مُحدَّدة من خوارزميات تعلُّم الآلة. إن بود وحدات مُعالجة التنسور الذي من إصدار ٢٠١٨ يُعالج ما يقرُب من ١٠^{١٧} عملية حسابيةٍ في الثانية؛ وهو تقريبًا نفس الرقم الذي يُعالجه كمبيوتر «ساميت»، لكنَّه يستهلك طاقةً أقل بقُرابة مائة مرة، كما أنَّ حجمه أصغر بمائة مرة أيضًا. وحتى لو ظلَّت تقنية الرقاقات كما هي ولم تصْغُر حجمًا، فمثل هذه الآلات يُمكن بيُسرٍ وبساطةٍ أن تُبنى على مقياسٍ أكبر لتُوفِّر مقدارًا هائلًا من الطاقة الحوسبيَّة المُخصَّصة لنظم الذكاء الاصطناعي.

ما سبق نقرة والحوسبة الكمِّية نقرة أخرى. إن الحوسبة الكمِّية تستخدم الخصائص الغريبة للدَّوال الموجيَّة في ميكانيكا الكم لتُحقِّق نتائج مُبهرة؛ فبضعف المكونات المادية الكمِّية، يُمكنُك معالجة «أكثر من ضعفي» عمليات الحوسبة! بصُورةٍ عامةٍ، يسير الأمر كالتالي:[33] لنفترض جدلًا أنَّ بحوزتك جهازًا صغيرًا يُخزِّن بتًّا كمِّيًّا أو كيوبت. هذا البت الكمِّي له حالتان؛ ٠ أو ١. من وجهة نظر الفيزياء التقليدية، فهذا الجهاز عليه أن يكون في حالةٍ واحدةٍ فقط من الحالتَين، أما في فيزياء الكمِّ، فإنَّ «الدَّالة الموجيَّة» التي تحمل معلومات عن البت الكمِّي تُخبرنا أنَّه يكون في الحالتَين معًا. فإن كان لديك بتان كمِّيان، فهناك أربع حالات وصلٍ مُحتملة: ٠٠، و٠١، و١٠، و١١. وإذا كانت الدَّالة الموجيَّة متشابكةً على نحوٍ مُترابط عبر البتَّين الكميَّين؛ أي لا تُوجَد أي عمليات فيزيائية أخرى لتُفسد هذا الترابط المُتناغم، حينها يكون البتَّان الكميان موجودَين في الحالات الأربع جميعها في الوقت نفسه. فضلًا عن ذلك، إذا كان البتَّان الكميَّان مُتصلَين في دائرة كمية تقوم ببعض العمليات الحسابية، فإن تلك العمليات الحسابية تُعالج في أربع الحالات في الوقت ذاته. أما إن كانت ثلاثة بتات كمية، فسيكون لديك ثماني حالاتٍ تُعالج في الوقت نفسه، وهكذا دواليك. ولكن هناك بعض القيود المادية لهذه العملية، بحيث إن مقدار العمل الناتج يكون أقل من المقدار الأُسِّي لعدد البتات الكمِّية،[34] ومع هذا فنحن نعلم علم اليقين أنَّ هناك مشاكل مُهمَّة ستتعامل معها الحوسبة الكمِّية بكفاءةٍ أعلى من نظيرتها التقليدية.

في عام ٢٠١٩، شهدنا بعض النماذج التجريبية لمُعالجاتٍ كمِّية صغيرة تحتوي على بضع عشراتٍ فقط من البتات الكمِّية، لكن حتى الآن لا تُوجَد أي مهامَّ حوسبيَّةٍ ذات أهميةٍ يتفوَّق المُعالج الكمِّي في سرعة أدائها على الكمبيوتر التقليدي. والعقبة الرئيسية

أمامنا تكمُن في إزالة التَّرابط الكمِّي؛ وإزالة التَّرابط الكمِّي هذا يحدُث عندما تُفسد بعض العمليات كالضَّوضاء الحرارية ترابطَ الدَّالة الموجيَّة ذات البتات الكمِّية المتعدِّدة. لكنَّ علماء فيزياء الكمِّ يأمُلون أن تُحلَّ هذه العقبة بدمج مجموعة دوائر مُصحِّحةٍ للأخطاء تكتشف سريعًا أي خطأ يحدُث في الحساب فتُصحِّحه بما يُشبه عملية التَّصويت. ولكن للأسف، تحتاج النظم المُصحِّحة للأخطاء إلى عددٍ أكبر من البتَّات الكمِّية لتعمل: ففي حين أنَّ جهازًا كمِّيًّا يحتوي على بضع مئاتٍ من البتات الكمِّية المثالية سيكون ذا قوةٍ هائلةٍ إذا ما قُورن بأجهزة الكمبيوتر التقليدية المعاصرة، لكن لنُدرك حقًّا حجم تلك القوة الجبارة، سنحتاج على الأرجح إلى بضعة ملايين من البتات الكمِّية المُصححة للأخطاء. والانتقال من بضع عشراتٍ من البتات الكمِّية إلى بضعة ملايين منها سيستغرق سنين عديدة ليتم، وحتى إذا ما وصلنا إلى تلك النقطة أخيرًا، حينها ستتغيَّر أفكارنا حول ماهية ما نستطيع تحقيقه باستخدام قوة الحوسبة المُطلقة هذه تغيُّرًا ثوريًّا.[35] فعوضًا عن انتظار اكتشافاتٍ تصوُّريَّة حقيقية في مجال الذكاء الاصطناعي، قد نتمكَّن من الاستعانة بطاقة الحوسبة الكمِّية الخارقة لنجتاز بعض العقبات التي تُواجه الخوارزميات «غير الذَّكية» الحالية.

(٢-١) حدود الحوسبة

حتى في خمسينيات القرن الماضي، كانت أجهزة الكمبيوتر تُلقَّب في الصُّحُف الشعبية بـ «العُقُول الخارقة» التي تعمل «أسرع من عقل أينشتين». ولكن ماذا عن اليوم؟ أيُمكننا أخيرًا أن نقول إنها تُضاهي في قوتها قوة العقل البشري؟ الإجابة هي لا! فالتَّركيز على قوة الحوسبة الهائلة وحدَها يحيد بنا عن الصَّواب تمامًا؛ فالسرعة بمُفردها لن تمنحنا ذكاءً اصطناعيًّا. إن تشغيل خوارزميَّة رديئة التَّصميم على كمبيوتر سريع لن يُحسِّن من أدائها، بل يعني فقط أنَّك ستحصُل على الإجابة الخطأ في وقتٍ أسرع. (وكُلَّما زاد حجم البيانات، زادت احتمالية الإجابات الخطأ!) كانت الغاية الرئيسية من الآلات السريعة، ولا تزال، هي اختصار وقت التَّجارب لتُنجز الأبحاثُ أسرع. إذن المكونات المادية ليست هي ما تكبح مسيرة الذكاء الاصطناعي، بل النُّظُم البرمجيَّة. فحتى الآن، نحن لا ندري كيف نجعل من آلةٍ ما كيانًا ذكيًّا حقًّا، حتى ولو كانت تلك الآلة بحجم الكون كله.

لكن لنفترض جدلًا أنَّنا نجحنا في تطوير النُّظُم البرمجيَّة المناسبة لبناء الذكاء الاصطناعي. هل تُوجَد أي حدودٍ فيزيائيةٍ لن تتخطاها قوة أجهزة الكمبيوتر؟ وهل

ستمنعُنا تلك الحدود من تملُّك ما يكفي من الطاقة الحوسبيَّة لصُنع ذكاءٍ اصطناعي حقيقي؟ والإجابة على هذين السؤالين هي نعم، هناك حدود فيزيائية ولكن لن تمنعنا ولا تُوجَد ولو ذرَّة من شكٍّ في ذلك. أقدم سيث لويد؛ الفيزيائي بمعهد ماساتشوستس للتقنية، على تقدير حدود كمبيوتر بحجم كمبيوتر محمول استنادًا على اعتباراتٍ من نظريتَي الكمِّ والقُصُور الحراري.[36] وكانت النَّتيجةُ صادمةً حتى إنها كانت تُدهش عالمًا مُخضرمًا ككارل سيجان؛ كانت النتيجة هي ١٠^٥١ عملية في الثَّانية الواحدة و١٠^٣٠ بايت من الذاكرة؛ بمعنًى آخر، أسرع بما يقرُب من مليار تريليون تريليون مرة من كمبيوتر «ساميت»، وأكبر بأربعة تريليونات مرة من ذاكرته؛ وقد أشرنا فيما سبق إلى أنَّ «ساميت» هذا يمتلك قوةً حوسبيَّةً تفوق العقل البشري. ولهذا عندما يسمع المرءُ منا أقاويل عن أنَّ العقل البشري يُمثِّل أعلى حدٍّ لما يُمكن تحقيقه فيزيائيًّا في هذا الكون الشاسع،[37] فعليه أن يُبادر على الأقلِّ بطلب توضيحٍ أكبر لهذا الادعاء.

إلى جانب الحدود التي تُمليها علينا الفيزياء، هناك حدود أخرى لقُدرة أجهزة الكمبيوتر نبعت من أبحاث علماء الكمبيوتر. آلان تورينج أثبت أنَّ بعض المشاكل بالنِّسبة إلى أيِّ كمبيوتر تكون «غير قابلة للحسم»؛ وبيان ذلك هو أن تكون المُشكلة مُعرَّفة تعريفًا دقيقًا وحلُّها معروف، لكن لا يُمكن أن تُوجد خوارزميَّة قادرة دائمًا على معرفة ذلك الحلِّ. وضرب مثلًا على ذلك سُمِّي فيما بعد بـ «مُعضلة التَّوقُّف»: هل تقدر أيُّ خوارزميَّةٍ على معرفة إذا ما كان برنامجٌ ما به «حلقة لا مُتناهية» تمنعُه من الاكتمال؟[38]

إثباتُ آلان تورينج أنه لا خوارزميَّة تقدر على حلِّ مُعضلة التَّوقُّف[39] هو إثبات في غاية الأهمية لأُسُس علم الرياضيات، لكنَّه لا علاقة له بمسألة ما إذا كان بإمكان أجهزة الكُمبيوتر أن تصير ذكيةً أم لا. وأحد الأسباب وراء هذا الادِّعاء هو أنَّ ذاك القُصُور الجوهري يبدو أنه ينطبق على العقل البشري. فإذا ما طلبت من أيِّ عقلٍ بشريٍّ أن يُحاكي نفسه مُحاكاةً دقيقةً، ثُمَّ يُحاكي تلك المُحاكاة، ثُمَّ تُحاكي هذه المُحاكاة الأخيرة نفسها وهكذا دواليك ... فحتمًا وبلا أدنى شكٍّ ستُواجه صعوباتٍ وعقباتٍ شتَّى. عن نفسي، لمْ يسبق لي القلق مُطلقًا حول قُصُوري فيما يتعلَّق بفعل هذا.

يبدُو إذن أن التركيز على المشكلات ذات القابلية للحسم لا يضع أي قيودٍ حقيقية للذكاء الاصطناعي. رغم ذلك، يتبيَّن لنا أنَّ كون مسألةٍ ما تقبل الحسم لا يعني أنَّ حسمها أمر هيِّن وسهل. يقضي علماء الكمبيوتر أوقاتًا طويلةً وهُم يُفكِّرون في مدى «تعقيد» المُشكلات؛ أي يتساءلون فيما بينهم عن كمية الحوسبة المطلوبة لحلِّ مُشكلةٍ ما بأكفأ

الطرق. وهاك مثالًا لمشكلة سهلة: أمامك قائمة بألف عدد، جِدِ العدد الأكبر فيما بينها. إن كنتَ ستتحقَّق من عددٍ واحدٍ في الثانية، فحلُّ هذه المسألة سيستغرق ألف ثانيةٍ إذا اتَّبعت هذه الطريقة الواضحة المُتمثلة في أن تتحقَّق من الأعداد عددًا واحدًا في كل مرة مع تذكُّر أيها أكبر قيمة. أهناك طريقة أسرع؟ لا، لأنَّه إذا تجاهلَتْ أيَّ طريقةٍ بعض الأعداد في القائمة، لا تدري لعل العدد الأكبر قيمةً يكون بين ما تجاهلته، وبهذا ستفشل في إيجاده. هكذا نجد أنَّ الوقت المُستغرق لإيجاد أكبر عنصرٍ في قائمةٍ ما يتناسب تناسبًا طرديًّا مع طُولها. قد تُعلِّق عالمة كمبيوتر على مثل هذه المسألة وتقول إنها مسألة ذات تعقيدٍ خطِّي؛ أي إن حلَّها سهل يسير. ثمَّ تَجِدُّ في البحث عن مسألةٍ أكثر أهمِّيةً وتشويقًا لتعمل على حلِّها.

إن ما يُثير اهتمام عالم كمبيوتر نظري هو حقيقة أنَّ الكثير من المشكلات في أسوأ الفُرُوض تبدو ذات صعوبةٍ «أُسِّيَّة».[40] وهذا يعني شيئين؛ الأول هو أنَّ جميع الخوارزميات التي نعرفها تتطلَّب زمنًا أُسِّيًّا — أي مقدارًا من الوقت يُمثِّلُ كأُسٍّ مرفوعٍ لحجم المُدخلات — لحلِّ بعض حالات المشاكل على الأقل؛ والثاني هو أنَّ علماء الكمبيوتر النَّظريِّين واثقون تمام الثِّقة أنَّ لا وجود لخوارزميات أكثر كفاءةً وفاعليَّة.

ونمو الصعوبة الأُسِّية يعني أنَّ المشكلات قد تُحَل نظريًّا؛ أي إنها بلا شكٍّ ذات قابليةٍ للحسم، لكنَّها تكون مُستعصيةً على الحلِّ عمليًّا أحيانًا؛ ونُسمِّي مثل هذا النوع من المشكلات بالمُشكلات «العسيرة». ومثالُ هذه المشكلات هو مُشكلة حسم ما إذا كانت خريطةٌ ما يُمكن أن تُلوَّن بثلاثة ألوانٍ فقط؛ بحيث لا يُلوَّن منطقتان متجاورتان فيها باللون نفسه أم لا. (من البديهي أنَّ تلوين الخريطة بأربعة ألوانٍ مُختلفةٍ هو حلٌّ مطرُوح في جميع الأحوال.) في تلك المشكلة، إذا كان عدد المناطق في الخريطة هو مليون، فقد نجد أنَّ بعض الحالات (بعضها وليس جميعها) يتطلَّب ما يُقارب $٢^{١٠٠٠}$ خطوة حوسبية لنصل إلى إجابة. وهذا الرَّقم يُساوي قُرابة $١٠^{٢٧٥}$ عام من الحوسبة إذا ما استخدمنا كمبيوتر «ساميت» الخارق، أو $١٠^{٢٤٢}$ عام فقط إذا استُخدِم كمبيوتر سيث لويد المحمول الذي يُلامس أقصى حدود القدرات الفيزيائية المُمكنة. هذا الرَّقم هائل لدرجة أنَّ عمر الكون الذي يُقدَّر بـ $١٠^{١٠}$ سنة تقريبًا لن يعدُو كونَه قطرة ماءٍ في محيطٍ واسع.

السؤال هنا: هل يجعلنا وجود مثل هذه المشكلات العسيرة نظُنُّ أن أجهزة الكمبيوتر لن يمكن أن تُضاهي البشر في الذكاء؟ لا؛ فنحن لا نفترض أنَّ البشر قادرون على حلِّ هذه

المشكلات العسيرة أيضًا. والحوسبة الكمِّية في هذه الحالات، سواء في الآلات أم الأدمغة، قد تُساعد قليلًا، لكن ليس بالقدْر الذي يُغيِّر من الناتج الأساسي.

والتَّعقيد يعني أنَّ مشكلة حسم القرارات في الحياة الواقعيَّة — كمشكلة اتِّخاذ قرارٍ بما ستفعله الآن في كل لحظةٍ من لحظات حياتك — هي مسألة غاية في الصعوبة، ولن يقدر البشر ولا أجهزة الكمبيوتر أبدًا في أيِّ وقتٍ قريبٍ أو بعيدٍ على إيجاد حلولٍ مثالية لها.

ونستشفُّ من ذلك استنتاجَين؛ أولهما أنَّنا نتوقع، في غالبيَّة الأوقات، أنَّ القرارات في الحياة الواقعية ستكون جيِّدةً على أحسن تقديرٍ، لكنَّها بعيدة كل البُعد عن المثالية؛ وثانيهما، أنَّنا نتوقَّع أنَّ جزءً كبيرًا من «البنية العقلية» للبشر والآلات؛ أي طريقة عمل عمليات اتِّخاذ القرارات، ستكون مُصمَّمة لتفادي التَّعقيد قدْر الإمكان؛ وهذا حتى نتمكَّن من أن نتوصَّل إلى تلك القرارات الجيدة رغم التَّعقيد الهائل في هذا العالم. وأخيرًا، نحن نتوقَّع أنَّ الاستنتاجَين السابقَين سيظلَّان حقيقةً مهما كان ذكاء وقوة الآلات التي قد تُصنع في المُستقبل؛ فالآلات قد تكون أكثر كفاءةً منَّا بكثيرٍ نحن البشر، لكنَّها ستكون بعيدة كلَّ البُعد عن العقلانية التامَّة.

(٣) أجهزة الكمبيوتر الذكية

أتاح تطوُّر المنطق على يد أرسطُو وغيره وضع أُسُسٍ دقيقة للتَّفكير العقلاني، ولكننا لا ندري إذا ما كان قد خطر على بال أرسطُو ذات مرةٍ أن يتفكَّر في احتمالية أن تُطبِّق الآلات تلك القواعد. في القرن الثالث عشر، اقترب رامون لول؛ الفيلسوف وزير النساء والمُتصوِّف الكتالوني الشهير، من هذه الفكرة وصنع بالفعل عجلاتٍ ورقيةً عليها رُمُوز منقوشة يستطيع من خلالها تكوين ودمج عباراتٍ منطقية. لكنَّ بليز باسكال عالم الرياضيات الفرنسي العظيم الذي عاش في القرن السابع عشر، كان أول من طوَّر آلةً حاسبةً ميكانيكية حقيقيَّة وعمليَّة. ومع أنَّها كانت لا تقدرُ إلا على جمع الأعداد أو طرحها، وكانت مُستخدَمة حصريًّا في مكتب أبيه لتحصيل الضرائب، فإنَّها أرشدت باسكال لكتابة ما يلي: «هذه الآلة الحسابية تُحدِث آثارًا تبدو أقرب إلى ما يُحدثه التَّفكير من كل السُّلوك الحيواني.»

حدثت في التقنية في القرن التاسع عشر طفرة هائلة عندما صمَّم تشارلز بابيج؛ عالم الرياضيات والمُخترع البريطاني، «المُحرِّك التحليلي»، الذي هو عبارة عن آلةٍ قابلةٍ للبرمجة

ومُتعدِّدة الأغراض بالمفهُوم الذي عرَّفه آلان تورينج لاحقًا. وقد ساعدَتْه في اختراعه ذاك آدا كونتيسة لوفليس، ابنة الشاعر الرومانسي والمستكشف اللورد بايرون. وبينما كان تشارلز بابيج يأمل في استخدام هذا المُحرِّك التحليلي في حساب بياناتٍ رياضيةٍ وفلكيَّةٍ دقيقة، فإن لوفليس انتبهت إلى القوة الحقيقية الكامنة في هذا المُحرِّك،[41] ووصفته في عام ١٨٤٢ باعتباره: «آلة تُفكِّر ... أو آلة لها القُدرة على الاستنتاج في كافَّة المجالات في هذا الكون.» وهكذا وُضعت المبادئ النَّظريَّة الأساسية لصُنع ذكاءٍ اصطناعي! ومن هذه النقطة من التاريخ، بلا شكٍّ كان ظهور الذكاء الاصطناعي مُجرَّد وقتٍ ليس إلا.

لسُوء الحظِّ، مرَّ وقت طويل لم يُبْن فيه المُحرِّك التحليلي أبدًا وباتت أفكار آدا لوفليس في طيِّ النِّسيان. ثمَّ جاءت أبحاثُ آلان تورينج النَّظريَّة عام ١٩٣٦ وما لحِقَها من زخم الحرب العالمية الثانية، فظهرت آلات الحوسبة العمومية على الساحة أخيرًا في أربعينيات القرن الماضي. ثمَّ ما لبثت أن ظهرت أفكار عن بناء ذكاءٍ اصطناعي في إثرها، وكانت ورقة آلان تورينج البحثيَّة التي نُشرت عام ١٩٥٠ تحت عنوان «الآلات الحاسوبية والذكاء»[42] هي أفضل ما كُتب من الأبحاث المُبكِّرة العديدة حول احتمالية بناء آلاتٍ ذكية. وقتها، كان المُشكِّكُون يجزمون بأنَّ الآلات من المُستحيل أن تفعل أيَّ شيءٍ يُمكن أن يجُول بخاطرك من أفعال البشر، لكنَّ آلان دحض تلك الشُّكُوك وفنَّدها. واقترح أيضًا اختبارًا عمليًّا للذكاء يُسمَّى «لُعبة المُحاكاة» والذي تطوَّر وصار في صُورةٍ أبسط ليُصبح ما يُعرف اليوم بـ «اختبار تورينج». وهذا الاختبار يقيس «سلوك» الآلة؛ وتحديدًا، يقيس مدى براعتها في خداع المُستجوب البشري بحيث تُقنعُه أنَّها أيضًا إنسان مثله.

إن لُعبة المُحاكاة لها دور مُحدَّد في ورقة آلان تورينج البحثيَّة؛ وهو أنها تجربة فكرية هدفها إخراس ألسنة المُشكِّكين الذين زعموا أنَّ الآلات لا يُمكنها أن تُفكِّر تفكيرًا سليمًا لأسبابٍ وجيهةٍ وبالقدْر المُلائم من الوعي. كان آلان يأمُل أن يُغيِّر اتجاه النقاش إلى مشكلة ما إذا كانت الآلات تستطيع أن تتصرَّف بطريقةٍ مُعيَّنة. وإذا تبيَّن أنَّها قادرة على ذلك؛ فهل تستطيع مثلًا أن تتناقش نقاشًا موزُونًا حول قصائد شكسبير ومعانيها؟ حينها لن يُمكن أن يدوم الشَّك في الذكاء الاصطناعي طويلًا. وخلافًا للتَّفاسير الشائعة، فإنِّي أشُكُّ أنَّ مثل هذا الاختبار كان يُقصَد به أن يُعرِّف الذكاء تعريفًا حقيقيًّا بمعنى أنَّ الآلة تكون ذكيةً فقط إذا اجتازت اختبار تورينج بنجاح. في الواقع، كتب آلان في ورقته البحثيَّة قائلًا: «ألا يُمكن للآلات أن تُنفِّذ شيئًا ما قد يبدُو كأنَّه تفكير في صُورته، لكنَّه في الحقيقة عملية مُختلفة تمامًا عن كيفية إعمال العقل لدينا نحن البشر؟» وسبب آخرُ

يدفعنا ألا نلتفت إلى ذلك الاختبار كتعريفٍ للذكاء الاصطناعي، وهو أنَّه لو كان تعريفًا لعُدَّ تعريفًا سيِّئًا جدًّا للعمل في ظلِّه. ولهذا السَّبب، لم يبذُل السَّواد الأعظم من باحثي الذكاء الاصطناعي أيَّ جُهدٍ يُذكر لاجتياز هذا الاختبار.

اختبار تورينج لا يُفيد الذكاء الاصطناعي؛ لأنَّه تعريف عام ومشرُوط للغاية؛ فهو يعتمد على خصائص العقل البشري الشديدة التَّعقيد والتي نجهل عنها أكثر بكثيرٍ مما نعلم، والمستمدة من التَّكوين البيولوجي والثقافي معًا. إنه لا سبيل إلى «تحليل» ذلك التَّعريف إلى مُكوِّنات أساسية يمكننا أن نسير عليها لنبني آلةً تستطيع أن تجتاز الاختبار. عوضًا عن ذلك، انكبَّ مجال الذكاء الاصطناعي على دراسة السلوك العقلاني كما وُضِّح آنفًا؛ أي تُعتبَر الآلة ذكيةً ما دام أنَّ فعالها يُتوقَّع منها على الأرجح أن تُحقِّق غايتها، مع أخذ مقدار إدراكها في الاعتبار.

استهل باحثو الذكاء الاصطناعي الأمر، كما فعل أرسطُو قبلهم، بالنظر إلى الغاية في عبارة «أن تُحقِّق غايتها»، باعتبارها هدفًا إما أن يُحقَّق أو لا. يُمكن أن تُوجد هذه الأهداف في عالم الألعاب مثلُ «أُحجية المربعات الخمسة عشر»، تلك التي يكون المطلوب فيها هو ترتيب مربعات الأرقام ترتيبًا تصاعديًّا من ١ إلى ١٥ في إطارٍ صغير مربَّع الشَّكل؛ أو قد تكون موجودة في بيئات مادية وواقعية. فمثلًا في أوائل سبعينيات القرن الماضي، كان الروبوت «شيكي» في معهد ستانفورد للأبحاث في كاليفورنيا كان يدفع المُكعَّبات الضخمة ليُشكِّل ترتيباتٍ مطلوبة، وكان الروبوت «فريدي» بجامعة إدنبرة يُجمِّع قاربًا خشبيًّا من أجزائه المُفكَّكة. كل هذا كان يُنجز باستخدام النظم المنطقية لحلِّ المُشكلات ونُظُم التخطيط لوضع وتنفيذ خُططٍ مضمونة لتحقيق الأهداف.[43]

وبحلول ثمانينيات القرن الماضي، كان من الجليِّ أنَّ التفكير المنطقي وحده لا يُمكن أن يفي بالغرض، وهذا لأنه كما أشرنا سابقًا، لا تُوجَد خطة «تضمن» لك الوصول إلى المطار. إن المنطق مبنيٌّ على اليقين والعالَم الذي نعيش فيه لا يُوجَد به شيء مؤكَّد. في تلك الأثناء، كان جوديا بيرل؛ عالم الكمبيوتر الأمريكي الإسرائيلي الذي فاز عام ٢٠١١ بجائزة آلان تورينج، مُنشغلًا بالعمل على طرائق للتفكير المنطقي غير المؤكَّد استنادًا إلى نظرية الاحتمالات.[44] وشيئًا فشيئًا تقبَّل باحثو الذكاء الاصطناعي أفكار بريل، وتبنَّوا آليات نظريتَي الاحتمالات والمنفعة؛ ومن ثمَّ تشابك علم الذكاء الاصطناعي مع غيره من العُلوم كعِلم الإحصاء ونظرية التَّحكُّم وعلم الاقتصاد وعلم أبحاث العمليات. وكان هذا التغيير علامةً فارقةً بدأ من بعدها ما يُسمِّيه بعض المُراقبين بـ «الذكاء الاصطناعي الحديث».

(١-٣) البيئات والكيانات

يتمحور الذكاء الاصطناعي الحديث حول مفهوم «الكيان الذكي»؛ وهو كيان يُلاحظ ويُدرك ويتصرَّف. وهو عملية تحدث بمرور الوقت بمعنى أنها تُحوِّل سلسلةً من المُدخلات المُدركة إلى سلسلةٍ من التصرفات. ولنضرب مثالًا على ذلك. لنفترض أنَّ الكيان الذكي هنا هو سيارة أجرةٍ ذاتية القيادة تُقلُّني إلى المطار. مُدخلات هذه السيارة قد تشمل ثماني آلات تصويرٍ آر جي بي تلتقط صورًا مُلوَّنة بمُعدَّل ثلاثين إطارًا في الثانية، وكل إطارٍ يحوي ما يُقارب ٧٫٥ ملايين بكسل، وكل منها له قيمة كثافة صورة في كلٍّ من قنوات الألوان الثلاثة؛ لينتج ما يربُو عن ٥ جيجابايتات في الثانية الواحدة. (يُعدُّ سيل البيانات المُتدفِّق من شبكية العين البشرية من خلال مائتي مليون مُستقبلٍ ضوئي بها أكبر حجمًا، وهذا جزئيًّا يُفسِّر لماذا تشغل حاسَّة البصر هذا الجزء الكبير من الدماغ البشري.) كما تشمل مُدخلات سيارة الأجرة أيضًا بياناتٍ من مقياس تسارُع بمُعدَّل مائة مرةٍ في الثانية الواحدة، جنبًا إلى جنبٍ مع بيانات نظام تحديد المواقع العالمي. يُحوِّل هذا السَّيل الهائل من البيانات الخام عبر المليارات من الترانزستورات (أو العُصبُونات) ذات القوة الحوسبية الجبارة إلى قيادةٍ سلسةٍ وفعالة. أما تصرفات سيارة الأجرة فتشمل الإشارات الإلكترونية المُرسلة إلى عجلة القيادة والمكابح ودوَّاسة الوقود بمُعدَّل عشرين مرةً بالثانية الواحدة. (جُلُّ هذه الدَّوامة من التصرفات المُتلاحِقة تتم على نحوٍ غير واعٍ بالنسبة إلى سائقٍ بشريٍّ مُحنَّك، ولا يعي الواحد منا إلا ما يُريد أن يتَّخذه من قراراتٍ مثل تخطِّي الشاحنة البطيئة التي تسير أمامه أو التَّوقف للتَّزُّود بالوقود، أما عيناه وعقله وأعصابه وعضلاته فهي تعمل معًا لتنفيذ بقية المهام.) إذا نظرنا إلى برنامجٍ للُعبة الشطرنج، فإن مُدخلاته تتلخَّص غالبًا في دقات الساعة التي تُشير إلى الوقت المُتاح لتنفيذ الحركة، بالإضافة إلى حركة خصمه على الرقعة ووضعها الجديد. أما التصرفات فهي إما أنَّه ساكن لا يفعل أي شيءٍ بينما يُفكِّر، وإما يختار حركته الجديدة من آنٍ لآخر ثم يُنبِّه الخصم. أما إذا تأمَّلنا مساعدًا رقميًّا شخصيًّا (بي دي إيه) مثل «سيري» أو «كورتانا»، فإن المُدخلات تتضمَّن أكثر من الإشارات الصوتية عبر الميكروفون (بمُعدَّل عينات يُساوي ثمانية وأربعين ألف مرةٍ في الثانية الواحدة) ومُدخلات من الشاشة اللمسيَّة، لتشمل أيضًا محتوى أي صفحةٍ من صفحات الإنترنت يزورها. بينما التصرفات تتضمن التَّحدُّث وعرض المعلومات على الشاشة.

تتوقَّف الطريقة التي نبني بها الكيانات الذكية على طبيعة المشكلة التي نواجهها. ومن ثمَّ، هذا يعتمد على ثلاثة عوامل؛ الأول: طبيعة البيئة التي سيعمل فيها هذا الكيان؛ فرُقعة الشطرنج بيئة مختلفة تمامًا عن أحد الطرق السريعة المزدحمة أو هاتفٍ جوال. أما العامل الثاني فهو المُلاحظات والتَّصرفات التي تربط الكيان بالبيئة، ومثال ذلك هو أنَّ «سيري» قد يكون لديه وصول لكاميرا الهاتف ليرى ما حوله أو لا. والعامل الثالث هو الغاية من الكيان؛ فتعليم الخصم أن يُطوِّر من مهاراته في الشطرنج مهمة مختلفة تمامًا عن تعليمه أن يفوز بالمُباراة.

ولنضرب مثالًا واحدًا فقط لتوضيح كيف يعتمد تصميم الكيان الذكي على تلك العوامل الثلاثة. إذا كانت الغاية هي الفوز بالمباراة، فإن أيَّ برنامجٍ مُصمَّم ليلعب الشطرنْج لا حاجة له أن يتذكر التحركات الماضية على الرُّقعة، بل يحتاج فقط إلى التَّفكير في وضعها الحالي.[45] على الجانب الآخر، يحتاج البرنامج الذي مُهمته تعليم الشطرنج أن يُحدِّث منهجه باستمرارٍ استنادًا على ما مضى من تحركاتٍ ليضُمَّ الجوانب التي استوعبها المُتعلم من قواعد الشطرنج وتلك التي لم يستوعبها بعدُ حتى يقدر على تقديم إرشادات مفيدة للمتعلم. بعبارة أخرى، بالنسبة إلى البرنامج الذي يُعلِّم الشطرنج يُعدُّ عقل المتعلم جزءًا ذا صلةٍ بالبيئة التي يعمل فيها البرنامج. وزد على ذلك أنَّه جزء لا يُمكن ملاحظته مباشرةً، على عكس الرُّقعة التي يراها أمامه مباشرةً.

إذا نظرنا إلى خصائص المشاكل التي تُؤثِّر على كيفية تصميم كيان ذكي، فسنجدُها تتضمَّن على الأقل ما يلي:[46]

- هل البيئة المحيطة يُمكن مُلاحظتها ملاحظةً كاملةً (كما في الشطرنج، حيث المُدخلات توفِّر وصولًا مباشرًا لجميع جوانب الوضع الحالي للبيئة المُحيطة ذات الصلة)؛ أم ملاحظةً جزئيةً (كما في قيادة السيارة حيث مجال رؤية السائق محدود ولا يُمكنه رؤية ما بداخل المركبات الأخرى ونوايا السائقين الآخرين مُبهمة؟)
- هل البيئة المحيطة والتصرفات مُنفصلتان إحداهما عن الأخرى (كما في الشطرنج)، أم مُتَّصلتان اتصالًا فعَّالًا (كما في قيادة السيارة)؟
- هل البيئة تضُمُّ كيانات أخرى (كما في الشطرنْج وقيادة السيارة)، أم لا (كما هو الحال أثناء إيجاد أقصر الطُّرُق إلى مكانٍ ما عبر الخريطة)؟

- بالاستناد إلى ما تنُصُّ عليه «قواعد» البيئة أو «قوانين الفيزياء» فيها، هل نتائج التصرفات قابلة للتَّوقُّع (كما في الشطرنْج)، أم لا يُمكن توقُّعها (كما في حركة المُرور وحالة الطقس)، وهل تلك القواعد التي استندنا عليها معلومة أم مجهولة؟
- هل البيئة تتغيَّر ديناميكيًّا فيكون الوقت المُتاح لاتخاذ القرار محدودًا (كما في قيادة السيارة)، أم لا (كما في اختيار الخُطَّة الضريبية المُثلى)؟
- ما مدى الإطار الزمني الذي تُقاس عليه جودة القرار المُتَّخذ وفقًا للغاية المُحدَّدة؟ هذا الإطار الزمني قد يكون قصيرًا جدًّا (كما في الضغط على مكابح السيارة في حالةٍ طارئة)، أو مُتوسط الطول (كما في لعبة الشطرنْج التي يصلُ عدد حركات المباراة فيها إلى حوالي مائة حركة)، أو طويلًا جدًّا (كما في رحلتي إلى المطار، والتي قد تتطلَّب مئات الآلاف من دورات اتخاذ القرار إذا افترضْنا أنَّ السائق يأخُذ مائة قرارٍ في الثانية الواحدة).

يُمكن للمرء منَّا أن يتخيل الكم المُحيِّر الذي تُثيره تلك الخصائص من مشكلاتٍ بأنواعٍ شتَّى. فإذا ما ضربنا بعض تلك الاختيارات ببعضٍ فسنحصُل على ١٩٢ نوعًا، ويُمكن لنا أن نجد مثالًا واقعيًّا لكل نوعٍ من تلك الأنواع. إن بعضها يُدرَس بطبيعة الحال في مجالاتٍ أخرى غير مجال الذكاء الاصطناعي؛ فمثلًا، تصميم نظام طيارٍ آليٍّ يُحافظ على مستوى تحليقٍ أفقي يُعدُّ مشكلة ديناميكية ومتَّصلةً وذات إطارٍ زمني قصير، وغالبًا ما تُدرَس في مجال نظرية التَّحكم.

من الواضح أن بعض المشكلات أسهل من غيرها. لقد أحرز مجال الذكاء تقدمًا كبيرًا في مشاكل كألعاب الطاولة والأحاجي التي تكون قابلة للملاحظة، ومنفصلة البيئة، ومحددة، ولها قواعد معلُومة سلفًا. بالنسبة لأنواع المشاكل الأسهل، فقد طوَّر باحثو الذكاء الاصطناعي خوارزميات ناجحةً وعامةً إلى حدٍّ ما، وكوَّنُوا عنها فهمًا نظريًّا مُتماسكًا، حتى إن الآلات غالبًا ما تتخطى البشر وتتفوَّق عليهم في الأداء في هذا النوع من المشاكل. ونحن نُطلق على خوارزميةٍ ما أنها خوارزمية عامة لأنَّنا نملك أدلةً رياضيةً على أنَّها تُعطي نتائج مثالية أو قريبة منها إذا ما طُبِّقت على فئةٍ كاملةٍ من المشاكل في ظل تعقيدٍ حوسبي مقبول، ولأنها تعمل جيدًا عمليًّا حين تُطبَّق دون الحاجة إلى تعديلاتٍ مُخصَّصة لكل مشكلةٍ على حدة.

أما ألعاب الفيديو مثل لُعبة «ستاركرافت»، فإنها تُعدُّ أصعبَ قليلًا من ألعاب الطاولة؛ فألعاب الفيديو بها المئات من العناصر المُتحرِّكة وأُطُر زمنية تشتمِل على الآلاف

من الخطوات، كما أنَّ الرُّقعة مرئية جزئيًّا في أي وقتٍ من الأوقات. في كل نقطة، يُمكن أن تصل الخيارات أمام اللاعب إلى ما لا يقلُّ عن $١٠^{٥٠}$ حركة، مقارنة بما يُقارب $١٠^{٢}$ في لُعبة مثل «جو».[47] ولكن على الجانب الآخر، القواعد معلُومة وبيئتها منفصلة بها أنواع محدودة من العناصر. بحلول عام ٢٠١٩، أصبحت الآلات تُحاكي في مهاراتها بعض أفضل لاعبي «ستاركرافت»، لكنها ليست مُستعدَّةً بعدُ لتُواجه أمهر اللاعبين البشريين على الإطلاق.[48] ما يُهمُّنا الإشارة إليه هنا هو أننا بذلنا مجهودًا كبيرًا ومُركَّزًا على تلك المُشكلة بعينها لنُحرز هذا التَّقدم؛ فالخوارزميات العامة لم تصل بعدُ إلى مرحلةٍ تُطبَّق فيها على لُعبة «ستاركرافت».

أما إذا ما نظرنا إلى مشاكل مثل إدارة حكومةٍ ما أو تدريس البيولوجيا الجزيئية، فسنجدها أصعبَ صعوبةً بالغة عما سبق. فتلك مشاكل ذات بيئاتٍ مُعقدةٍ غالبًا ما تكون غير قابلة للمُلاحظة (حالة دولة بأكملها أو حالة عقل طالب)، وتحتوي على عناصر وأنواعٍ؛ عناصر أكثر بكثير، ولن تجد تعريفاتٍ واضحةً عن ماهية التصرفات، كما أنَّ أغلب القواعد مجهولة، زد على ذلك وجود الكثير من الشَّك وعدم اليقين، وأطرٍ زمنية طويلة جدًّا. نحن نمتلك الأفكار والأدوات الجاهزة التي نتعامل بها مع كل خاصيةٍ من تلك الخصائص على حدة، لكن حتى الآن لا تُوجَد طرق عامة تتماشى مع جميع الخصائص معًا في وقتٍ واحد. عندما نبني نظم ذكاءٍ اصطناعي لحلِّ هذا النوع من المهام، فإن تلك النظم تتطلَّب كمًّا هائلًا من التَّصميم المُخصَّص، وغالبًا ما تكون هشةً للغاية.

إحراز التقدم فيما يتعلَّق بالتوصُّل إلى خوارزميات عامة يحدث عندما نبتكر طرائق فعالة تُستخدم لمُعالجة المشكلات الصَّعبة في فئةٍ ما، أو عندما نُصمِّم طرائق تتطلَّب افتراضاتٍ أقل وأسهل بحيث يمكن أن يُعمَّم تطبيقها على مشكلاتٍ أكثر. إن الذكاء الاصطناعي العام طريقة قابلة للتطبيق في جميع فئات المُشكلات، تعمل بفعاليةٍ عند تطبيقها على المشكلات الأصعب والأكثر تعقيدًا مع استخدام افتراضاتٍ قليلة جدًّا. وهذه هي الغاية الأسمى لأبحاث الذكاء الاصطناعي؛ ابتكار نظامٍ لا يحتاج إلى تصميمٍ مُخصَّصٍ لمُشكلةٍ بعينها ويمكنه أن يُدرِّس محاضرة في علم البيولوجيا الجزيئية أو يدير حكومة دولةٍ ما. إنه نظام يتعلَّم ما يحتاج إلى معرفته من خلال جميع المصادر المُتاحة له، ويطرح الأسئلة حين الحاجة ثمَّ يبدأ في وضع الخطط الفعالة وتنفيذها.

مثل هذا النِّظام العام ليس موجودًا في الوقت الحالي، لكنَّنا نقترب منه شيئًا فشيئًا. وقد تتفاجأ حين تعلم أنَّ مقدارًا كبيرًا من هذا التَّقدُّم تجاه ذكاءٍ اصطناعي عامٍّ يُنسَب

إلى أبحاثٍ لا تتمحور حول بناء نظم ذكاء اصطناعي عامة. هذا التَّقدم يُنسب إلى أبحاثٍ في مجال «الذكاء الاصطناعي المحدود» أو «الذكاء الاصطناعي الخاص» والذي يعني نظم ذكاءٍ اصطناعي لطيفة وآمنة ومُملَّة صُمِّمت لحل مشكلات بعينها مثل لعب لُعبة جو أو التَّعرف على الأرقام المكتوبة يدويًّا. إن الأبحاث في هذا النوع من الذكاء الاصطناعي يُظَنُّ عادة أنها لا تُمثِّل أي خطرٍ لأنها مُصمَّمة لغرضٍ بعينه، ولا علاقة لها بالذكاء الاصطناعي العام.

إنَّ هذا الاعتقاد إنما هو ناجم عن سوء فهمٍ لنوعية العمل الذي تنطوي عليه تلك النظم. في الحقيقة، غالبًا ما تُعطي أبحاث الذكاء الاصطناعي المحدود دفعةً باتجاه الذكاء الاصطناعي العام، وخصوصًا عندما تُجرى على يد باحثين يتصدون للمشاكل التي تتخطَّى حدود قدرات الخوارزميات العامَّة الحالية، وطريقتهم في حل المشكلات ليست مجرَّد حشو ما يلزم من شفراتٍ خاصة لمُحاكاة تصرُّفات شخصٍ ذكي إذا وُضع في هذا الموقف أو في ذاك، إنما محاولات لغرس قدرةٍ ما في الآلات تُصبح بعدَها قادرةً بنفسها على إيجاد حلول للمشاكل التي تُواجهها.

ومثال ذلك هو نجاح فريق «ألفا جو» بشركة ديب مايند التابعة لجوجل في تصميم برنامجهم الذي سحق أبطال العالم في لُعبة جو، حيث حقَّقوا هذا الإنجاز دون بذل الجهد في تعليم البرنامج اللعبة ذاتها. وما أعنيه بذلك هو أنَّهم لم يكتُبوا مجموعة كبيرة من السطور البرمجية الخاصة بلُعبة جو والتي تُحدِّد ما الذي يجب على البرنامج فعله في المواقف المختلفة أثناء اللعب. ولم يضعوا إجراءات لاتخاذ القرارات خاصة بلُعبة جو دون غيرها. عوضًا عن ذلك، ما فعلوه هو أنَّهم طوَّروا أسلوبَين من الأساليب العامة إلى حدٍّ ما تطويرًا كافيًا للعب لُعبة جو بمهاراتٍ خارقة تفوق قدرة البشر، وهذان الأسلوبان هما أسلوبا البحث الاستباقي المُعِين على اتِّخاذ القرارات، وأسلوب التَّعلُّم المُعزز لتعلُّم كيفية تقييم الأوضاع. هذا التَّطوير قابل للتَّطبيق على مشكلاتٍ أخرى كثيرة، بما في ذلك المشكلات في مجالاتٍ بعيدةٍ كل البُعد كمجال صناعة الروبوت. ولمزيد من إبراز النجاح، دعني أُخبرك أنَّ إصدارًا من برنامج «ألفا جو» يُسمَّى «ألفا زيرو» تعلَّم مؤخَّرًا كيف يهزم إصدار «ألفا جو» في لُعبة جو، كما تعلَّم كيف يهزم «ستوك فيش» (وهو أفضل برنامج يلعب الشطرنج في العالم وبمهاراتٍ خارقة تفوق قدرات البشر) و«إلمو» (وهو أفضل برنامج للُعبة الشطرنج الياباني «الشُّوجي» والذي تفوق مهاراته أي كائن بشري). كل هذه الانتصارات حقَّقها برنامج «ألفا زيرو» في يومٍ واحد لا غير.[49]

كما كان هناك تقدُّم مُعتبر باتِّجاه الذكاء الاصطناعي العام، نابع من الأبحاث التي أُجريت في تسعينيات القرن الماضي للتَّعرف على الأرقام المكتوبة يدويًّا. لم يكتب فريق يان ليكن بمُختبرات شركة إيه تي آند تي أيَّ خوارزمياتٍ خاصة للتعرُّف على الرقم ٨ عن طريق البحث عن الخطوط المُنحنية والحلقات، بل طوَّروا خوارزميات للتَّعلُّم بالشَّبكات العصبونية موجودة بالفعل ليُنتجوا «الشبكات العُصبُونية الالتفافيَّة»، التي أظهرت بدورها نجاحًا في التَّعرُّف على الرموز بعد تدريبٍ مناسب على الأمثلة ذات الصلة. تلك الخوارزميات نفسها يُمكن أن تتعلَّم كيف تتعرَّف على الحروف والأشكال وعلامات التَّوقُّف والكلاب والقطط وسيارات الشُّرطة. وتحت مُسمَّى «التَّعلُّم المُتعمِّق»، أحدثوا ثورةً في مجالَي التَّعرُّف على الكلام وتمييز العناصر المرئية اللذَين يُعدَّان حجر الأساس في برنامج «ألفا زيرو» ومعظم المشاريع المُعاصرة لبناء سياراتٍ ذاتية القيادة.

وإذا ما أمعنت النظر في الأمر، فإنَّك لن تجد غرابةً في إدراك أنَّ التَّقدُّم نحو خوارزميات ذكاءٍ اصطناعي عامة سيحدث عبر مشاريع الذكاء الاصطناعي المحدود التي تُنفِّذ مهامَّ مُحدَّدة؛ فتلك المهام هي التي تجعل باحثي الذكاء الاصطناعي مُنهمِكين في البحث والتطوير. (هناك سبب يجعل الناس لا يقولون: «التَّحديق خارج النافذة هو أمُّ الاختراع».) في الوقت نفسه، من المُهم أن نفهم مدى التَّقدم الذي أُحرز وأين هي حُدود ذلك التَّقدم. عندما هزم برنامج «ألفا جو» لي سيدول ثمَّ لاحقًا سحق جميع عمالقة لُعبة جو الآخرين، افترض العديد من الناس أنَّ هذا الانتصار هو بداية النِّهاية، وما هي إلا مسألة وقتٍ ليس إلا حتى نرى الذكاء الاصطناعي يُسيطر على العالم؛ كُل هذا لأنَّ إحدى الآلات قد تعلَّمت من الصِّفر هزيمة مُنافسيها من البشر في مُهمَّةٍ يُعرف عنها صعوبتها الشديدة حتى بالنسبة إلى أكثر البشر فطنةً ودهاءً. إن بعض المُشكِّكين في موضوع سيطرة الذكاء الاصطناعي قد زالت شُكوكهم واقتنعوا عندما فاز برنامج «ألفا زيرو» في الشطرنج والشُّوجي، بالإضافة إلى لُعبة جو. لكن يجدُر الإشارة هنا إلى أنَّ برنامج «ألفا زيرو» له قيود صارمة؛ فهو لا يعمل إلا في فئة الألعاب الثنائية اللاعبين، ذات القواعد المعلُومة سلفًا، والتي تكون غير قابلة للملاحظة، وفي بيئةٍ مُنفصلة. ببساطةٍ، مثل هذا الأسلوب لن ينجح مُطلقًا في قيادة السيارات أو التَّدريس أو تولِّي قيادة حكومة دولةٍ ما، أو السيطرة على العالم.

تلك القُيُود الشديدة على كفاءة الآلات تعني أنَّه عندما يتحدث الناس عن ازدياد «مُعدَّل ذكاء الآلات» ازديادًا فائقًا يُنذِر بأنَّه سيتخطَّى معدلات الذكاء البشرية، فإن

كلامهم هذا ما هو إلا لغط لا قيمة له مُطلقًا. ونحن إذْ نقول إن مفهوم مُعدَّل الذكاء قد يبدو منطقيًّا إذا ما طُبِّق على البشر، فهذا لأنَّ القُدرات البشرية عادةً ما يترابط بعضها مع بعض في نطاقٍ كبير من الأنشطة المعرفية. ومُحاولة تعيين معدل ذكاءٍ للآلات تُشبه محاولة أن تأتي بحيوانٍ يمشي على أربعٍ وتضعه في مُنافسةٍ مع البشر في مُسابقة العُشاري الخاصة بألعاب القوى. لا أحد يُنكر أنَّ الخيل تستطيع أن ترمح رمحًا سريعًا وأن تقفز عاليًا، لكنَّها ستُواجه الكثير من الصُّعوبات في رياضتَي القفز بالزَّانة ورمي القُرص.

(٣-٢) الغايات والنموذج القياسي

بالنظر إلى أي كيانٍ ذكي من الخارج، ما يُهمُّنا هو سلسلة التَّصرفات التي يتَّخذها استنادًا إلى سيل المُدخلات التي يستقبلها. أما من الداخل، فالتَّصرفات يجب أن تُتَّخذ بمعرفة ما يُطلق عليه «برنامج كيان». إن جاز القول إن البشر يُولدون ببرنامج كيان واحد، ثمَّ يبدأ هذا البرنامج بالتَّعلُّم بمرور الوقت ليتَّخذ تصرُّفاتٍ ناجحةً بقدرٍ معقولٍ في عددٍ ضخمٍ من المهام. حتى الآن، ليس هو الحال بالنسبة للذكاء الاصطناعي؛ فنحن لا نعرف كيفيةَ بناء برنامج ذكاءٍ اصطناعي عامٍّ يفعل كل شيءٍ، لهذا وعوضًا عن ذلك، نبني أنواعًا مختلفةً من برامج الكيان التي يختصُّ كلُّ واحدٍ منها بنوعٍ مختلف من المشاكل. وأظُنُّ أني بحاجةٍ إلى شرح ولو جزءٍ يسيرٍ من كيفية عمل تلك البرامج المختلفة؛ وفي الملاحق في نهاية الكتاب سيجدُ القارئ المُهتم شرحًا مُستفيضًا لهذه النُّقطة. (وُضعت الإشارات إلى ملاحق بعينها كحروفٍ صغيرة بين قوسين أعلى الكلام هكذا«أ» وهكذا«د».) وسأُركِّز في هذه الجُزئية تركيزًا أساسيًّا على كيفية تمثيل النمُوذج القياسي في مُختلف أنواع الكيانات؛ بعبارةٍ أخرى، كيفية تحديد الغاية ونقلها إلى الكيان.

أبسط طرائق نقل الغاية هي أن تُنقل في صيغة «هدف». عندما تستقلُّ سيارتك الذاتية القيادة ثمَّ تضغط على أيقونة «البيت» الظَّاهرة على الشاشة، تستقبل السيارة هذا كغايةٍ يجب بُلوغها ثمَّ تشرع في انتقاء الطَّريق وبدء الرحلة. بعد ذلك إما أن يُحقق العالم الواقعي الهدف المرجو (أجل، وصلت إلى البيت) أو لا يُطابقه (لا، أنا لا أسكن في مطار سان فرانسيسكو). في الحقبة الكلاسيكية لأبحاث الذكاء الاصطناعي وقبل أن تُصبح مشكلة الارتياب وعدم اليقين هي المشكلة الرئيسية في ثمانينيات القرن الماضي، كانت غالبية أبحاث الذكاء الاصطناعي تفترض عالمًا محددًا وقابلًا للملاحظة بالكامل،

وتُعدُّ فيه الأهداف طريقةً منطقيةً لتحديد الغايات. أحيانًا تكون هناك أيضًا «دالة تكلفة» لتقييم الحُلُول، وهكذا يكون الحل المثالي هو الذي يُقلِّل من التَّكلفة الإجمالية ويصل إلى الهدف في الوقت ذاته. إذا طبَّقنا هذا على السيارة، فلرُبَّما يكون هذا مدمجًا فيها تتخذه تلقائيًّا — ربما تكون تكلفة طريقٍ ما هي إلا مُحصِّلة ثابتة لمجموع الوقت المُستغرق والوقود المُستهلك معًا، أو قد يكون للراكب البشري الخيار في تحديد أيهما سيُضحِّي به في سبيل الآخر.

والسَّبيل إلى تحقيق مثل تلك الغايات يكمُن في القدرة على «المُحاكاة الذهنية» لتأثيرات التصرُّفات المُحتملة والتي تُسمَّى أحيانًا بـ «البحث الاستباقي». إن سيارتك الذاتية القيادة مُزوَّدة بخريطةٍ مُدمجة؛ ولهذا فهي تعرف أنَّك إذا كنت في سان فرانسيسكو واتَّجهتَ شرقًا عبر جسر سان فرانسيسكو–أوكلاند فستصل إلى أوكلاند. وهكذا تجد الخوارزميات التي بُنيت في ستينيات القرن العشرين[50] طُرقًا مثاليةً فقط بالبحث الاستباقي وفحص العديد من تسلسلات التَّصرُّفات المُحتملة.«أ» تلك الخوارزميات تُشكِّل السَّواد الأعظم من البنية التَّحتية الحديثة؛ فهي لا تُعطي اتجاهات القيادة فقط، بل تُوفِّر حلولًا للسَّفر الجوي وتجميع الروبوتات والتَّخطيط العمراني والإدارة اللوجستيَّة للتوريدات. وبإجراء بعض التعديلات للتَّعامل مع السُّلوك المُفاجئ للخُصُوم، فإنَّ الفكرة ذاتها الخاصة بالبحث الاستباقي تُطبَّق في ألعابٍ مثل الشطرنج وجو وإكس–آو، حيث الهدف هو الفوز وفقًا لما يعنيه مفهوم الفوز في كل لُعبة.

تعمل خوارزميات البحث الاستباقي بكفاءةٍ لا مثيل لها في المهامِّ المُحدَّدة المكلفة بها، لكنَّها لا تتَّسم بالمُرُونة الكافية. فمثلًا، برنامج «ألفا جو» «يعرف» قواعد لُعبة جو فقط بصفتها تحوي روتينين فرعيَّين بُنِيا بلُغةٍ برمجيةٍ تقليديةٍ مثل «سي++»؛ روتينًا يُولِّد جميع التَّحرُّكات الجائزة المُحتملة، والآخر يُشفِّر الهدف ثمَّ يُقرِّر ما إذا كان الوضع الحالي فوزًا أم خسارة. وليلعب برنامج «ألفا جو» لُعبةً أخرى، على أحدٍ ما أن يُعيد كتابة تلك الشَّفرة المكتوبة بلُغة «سي++» بالكامل. علاوة على ذلك، إذا أوكل له هدف جديد — لنقُل مثلًا: زيارة الكوكب غير الشمسي الذي يدور حول نجم «القنطُور الأقرب» — ما سيفعله البرنامج حينها هو أنَّه سيسبُر أغوار المليارات من سلاسل تحرُّكات لُعبة جو بحثًا عن تسلسلٍ يُحقِّق الهدف الجديد، ولكن بلا طائل. فهو لا يُمكنه فحص شفرته التي بلُغة «سي++» والانتهاء إلى الحقيقة الواضحة؛ ألا وهي: لن تُجدي أي سلسلةٍ من

سلاسل التَّحرُّكات في لُعبة جو نفعًا في إيصاله إلى الكوكب المطلوب. فمعارف البرنامج بصفةٍ أساسيةٍ محبُوسة في داخل صندوقٍ أسود.

في عام ١٩٥٨ وبعد أن مرَّ عامان على برنامج دارتموث الصيفي الذي أرسى قواعد مجال الذكاء الاصطناعي، اقترح جون ماكارثي منهاجًا أعمَّ وأوسع بإمكانه فتح ذاك الصندوق الأسود، والذي تمثَّل في بناء برامج تفكير عامةٍ يُمكنُها أن تتشرَّب المعرفة في أيِّ مجالٍ كان، ثمَّ تُنعم النَّظر في تلك المعرفة لتُجيب عن أيِّ أسئلةٍ يُمكن الإجابة عليها.[51] وأخُصُّ بالذكر أحد أنواع التَّفكير الذي اقترحه أرسطُو وهو «التفكير العملي»: «إذا قمت بالأفعال «أ» و«ب» و«ج» ... فستُحقِّق الهدف «ز».» وهذا الهدف قد يكون أي شيء على الإطلاق؛ قد يكون مثلًا: رتِّب البيت قبل أن أصل، أو فُز في مباراة شطرنج دون أن تخسر أيًّا من الحصانَين، أو خفِّض من ضرائبي بنسبة ٥٠ بالمائة، أو زُر نجم «القنطُور الأقرب» وهلمَّ جرًّا. سُرعان ما أصبحت هذه الفئة الجديدة من البرامج التي اقترحها جون ماكارثي تُعرف باسم «النظم القائمة على المعرفة».[52]

ولنبني نظمًا قائمةً على المعرفة، علينا أن نُجيب على سؤالين لا ثالث لهما. الأول هو: كيف تُخزَّن المعرفة داخل كمبيوتر؟ أما الثاني؛ فكيف للكمبيوتر بعد إذنٍ أن يُفكِّر تفكيرًا صحيحًا استنادًا إلى تلك المعرفة ليصل في النِّهاية إلى استنتاجات جديدة؟ ولحُسن حظنا، أجاب فلاسفة اليونان القديمة، وخصوصًا أرسطُو، على تلك الأسئلة بإجاباتٍ أساسية قبل مجيء أجهزة الكمبيوتر إلى عالمنا بوقتٍ طويل. في الحقيقة، أجد جليًّا أن أرسطُو لو كان لديه كمبيوتر (وتيار كهربي أيضًا) لكان قد اشتغل كباحثٍ في مجال الذكاء الاصطناعي. وإجابة أرسطُو على هذَين السؤالين، كما أعاد طرحها جون ماكارثي، هي أن نستخدم المنطق الصُّوري«ب» كحجر أساسٍ للمعرفة والتَّفكير.

هناك نوعان من المنطق يُعدَّان مُهمَّين في علم الكمبيوتر. النَّوع الأول يُسمَّى «منطق القضايا» أو «المنطق البوليني»، وقد كان معروفًا عند اليونانيِّين القدماء، والفلاسفة الهنود، والصينيين القدماء. وهو يعتمد على بوابات «الاقتران» وبوابات «العاكس المنطقي» وغيرهما من البوابات التي تُشكِّل مجموعة الدوائر الكهربية في رقاقات الكمبيوتر. وإذا تمعنا في وحدة معالجةٍ مركزيةٍ حديثةٍ فإننا سنجدها، بالمعنى الحرفي للكلام، عبارةً عن تعبيرٍ رياضيٍ غاية في الطول — قد يحتاج لمئات الملايين من الصفحات — مكتوب بلُغة منطق القضايا. أما النوع الثاني، فهو ذاك النوع من المنطق الذي اقترح جون ماكارثي استخدامه في مجال الذكاء الاصطناعي ويُسمَّى «المنطق الإسنادي».«ب» وتُعدُّ لُغة المنطق

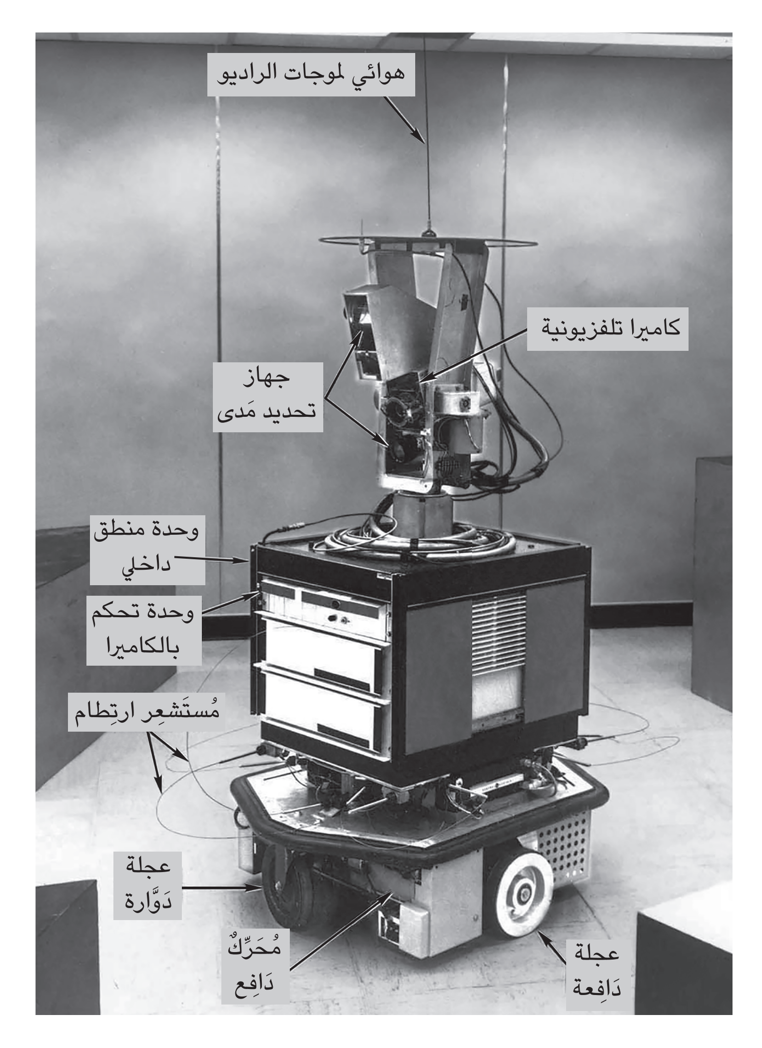
الإسنادي لُغةً ذات قدرةٍ تعبيرية أكبر بكثيرٍ من منطق القضايا، مما يعني أن هناك أشياء يمكننا التعبير عنها بسهولةٍ ويُسرٍ باستخدام المنطق الإسنادي والتي تكون شاقةً أو تكاد تكون أمرًا مستحيلًا إذا حاولنا كتابتها باستخدام منطق القضايا. ومثالُ ذلك هو أنَّ قواعد لُعبة جو تُكتب في نحو صفحةٍ واحدةٍ بلُغة المنطق الإسنادي، لكنَّها تأخذ قُرابة عدة ملايين صفحةٍ باستخدام منطق القضايا. بالمثل، يُمكننا التَّعبير بسهولةٍ باستخدام هذا النوع من المنطق عن معارف كثيرةٍ كقواعد الشطرنج، ومعنى المُواطنة البريطانية وقانون الضَّرائب والبيع والشراء، والتنقل والرَّسم والطَّهي وغيرها العديد من المناحي البديهيَّة في عالمنا.

من حيث المبدأ إذن، القُدرة على التَّفكير بالمنطق الإسنادي تجعلنا نقطع شوطًا كبيرًا باتِّجاه الذكاء الاصطناعي العام. في عام ١٩٣٠، نشر عالم المنطق النِّمساوي المُخضرم كورت جوديل بحثه الشهير «مبرهنة تمام المنطق الإسنادي»[53] الذي أثبت فيه أنَّ هناك خوارزمية ما تتَّسم بالخاصية التالية:[54]

> لأي «نوع من المعرفةٍ» أو لأي «سؤال» قابل للتعبير عنه بالمنطق الإسنادي، فإن الخوارزمية ستُخبرنا بالإجابة عن هذا السؤال إذا كانت الإجابة موجودة أصلًا.

وهذا ضمان مُذهل! إنه يعني أنَّنا، على سبيل المثال، يُمكن أن نُعلِّم النِّظام قواعد لُعبة جو وهو سيُخبرنا (إذا انتظرنا الوقت الكافي) إذا ما كانت هُناك حركة افتتاحيَّة تُمكِّن صاحبها من الفوز بالمُباراة. كما يُمكننا أن نُغذِّي النظام بالحقائق الجُغرافيَّة لمنطقةٍ ما وسيهدينا إلى طريق المطار. بل ويُمكننا أن نُعلِّمه الحقائق الهندسية وقوانين الحركة وماهية أدوات المطبخ، وسيُعلِّم ذلك النظامُ الروبوت كيفية ترتيب مائدة طعام العشاء. وعلى وجه العُمُوم، إذا ما أعطي الكيان أي هدفٍ قابلٍ للتحقيق ثمَّ غُذِّي بالمعلومات الكافية ليعرف نتائج تصرُّفاته وأثرها، فيُمكنه أن يستخدم الخوارزمية لوضع خُطَّةٍ ليُنفِّذها ليُحقِّق هذا الهدف.

علينا أن نُبيِّن أن كورت جوديل لم يُقدِّم أي خوارزمية، بل أثبت فقط أنَّ هناك واحدة موجودة. وفي بواكير ستينيات القرن الماضي، بدأت الخوارزميات الحقيقية للتَّفكير المنطقي بالظُّهور،[55] ولاح في الأفق حُلُم جون ماكارثي ببناء نظم ذات ذكاءٍ عامٍّ قائمة على المنطق، وبدا قريبًا من أن يُصبح حقيقة. وأول الغيث كان مشروع الروبوت المُتنقِّل المُسمَّى بـ «شيكي» والذي كان من تطوير معهد ستانفورد للأبحاث، والذي كان قائمًا على التَّفكير المنطقي (انظر الشَّكل ٢-٢). تلقَّى «شيكي» هدفًا ما من مُطوِّريه البشريين، فاستخدم



شكل ٢-٢: الروبوت «شيكي» عام ١٩٧٠ تقريبًا. في الخلفية تُوجد بعض الأشياء التي دفعها «شيكي» هنا وهُناك حتى تستقرَّ في أماكنها الصحيحة.

خوارزميات الإبصار لاستنتاج تأكيداتٍ منطقيةٍ لوصف الوضع الحالي، ثمَّ أجرى استدلالًا منطقيًّا لوضع خُطَّةٍ تضمن تحقيق الهدف ثمَّ بدأ بتنفيذها. كان الروبوت «شيكي» بمثابة

دليلٍ «حيٍّ» على أنَّ تحليل أرسطُو للمعرفة والسُّلوك البشريَّين كان تحليلًا صحيحًا جزئيًّا على الأقل.

لكن للأسف، كان ذاك التحليل الذي افترضه أرسطو (ومن بعده جون ماكارثي) بعيدًا كل البُعد عن كونه تحليلًا كامل الصحَّة لا تشُوبه شائبة. فالمُعضلة الأساسية هي الجهل، ولا أقصد هنا جهلًا عند أرسطو أو جون ماكارثي، بل الجهل في جنسنا البشري وفي الآلات أيضًا حاضرًا ومُستقبلًا. إن مقدارًا ضئيلًا جدًّا من معرفتنا هو ما يُمكن اعتباره معرفةً يقينية. وأخصُّ بالذكر هنا معرفتنا عن المُستقبل التي لا تكاد تُذكر. إن الجهل مُعضلة لا تُذلَّل لأيِّ نظامٍ منطقي صرف. فلو سألت مثلًا: «هل سأصل إلى المطار في الوقت المُحدَّد إذا غادرت المنزل ثلاث ساعاتٍ قبل موعد الرحلة؟» أو «هل يُمكنني أن أتملَّك منزلًا إذا اشتريتُ بطاقة يانصيب رابحة ثمَّ اشتريت المنزل بنُقُود الجائزة؟» هنا الإجابة الصحيحة لكِلا السؤالين هي: «لا أدري!» والسبب وراء ذلك هو أنَّ الإجابة عن أيٍّ من السؤالَين سواء بنعمٍ أم بلا، كلاهما احتمال منطقي صحيح. ومن الناحية العملية، فلا يُمكن للمرء أن يحظى بإجابةٍ يقينيةٍ عن أي سؤالٍ تجريبي إلا إذا كانت الإجابة معروفةً قبلًا.[56] ولحُسن الحظِّ، لا يُعدُّ اليقين ضرورةً لاتِّخاذ التَّصرُّفات؛ فكل ما نحتاج إلى معرفته هو أي التَّصرفات أفضل، لا أيها حتمًا سينجح.

وعدم اليقين هنا يعني أنَّ «الغاية التي جعلنا الآلة تسعى لتحقيقها» لا يُمكن في العُمُوم أن تكون هدفًا موصُوفًا بدقةٍ بحيث يتمُّ تحقيقه مهما كان الثمن. فلم يعُد هناك ما يُسمَّى بـ «سلسلة من التَّصرُّفات التي تُفضي بالهدف إلى تحقيقه»، وهذا لأنَّ أي سلسلةٍ من التَّصرفات سيكون لها العديد من النتائج المحتملة، والتي بعضها لن يُحقِّق الهدف المطلوب. وهنا نرى أنَّ أرجحيَّة النَّجاح مسألة مُهمَّة؛ فالمُغادرة إلى المطار قبل موعد الرحلة بثلاث ساعاتٍ «رُبَّما» يعني أنَّك لن تُفوِّت الطائرة، وشراء تذكرة يانصيب «رُبَّما» يعني أنَّك قد تربح ما يكفي من النقود لشراء منزلٍ جديد، وشتان بين «رُبَّما» الأولى و«رُبَّما» الثانية. فالأهداف لا يُمكن أن تُنقذ من الفشل بالبحث عن خُطط تُعزِّز من أرجحيَّة تحقيقها؛ فالخُطَّة التي تزيد من أرجحيَّة الوصول إلى المطار في الوقت المُحدَّد للَّحاق بالرحلة قد تجعلك تُغادر البيت عدة أيامٍ قبل الموعد وبرفقتك حُراس مُسلَّحين بينما ينتظرك عدد من وسائل النَّقل البديلة لنقلك في حال تعطَّلت سيارة الأجرة التي تستقلُّها، وما إلى ذلك من التَّجهيزات. لكن حتمًا، لا بُدَّ للمرء أن يضع في اعتباره التَّفضيلات النِّسبية للنتائج المُختلفة جنبًا إلى جنبٍ مع أرجحيَّة حدوثها.

إذن، يُمكننا أن نستعيض عن الأهداف باستخدام «دالَّة المنفعة» لوصف تفضيلات النتائج المُختلفة أو سلاسل الأوضاع. وغالبًا ما يُعبَّر عن منفعة سلسلةٍ ما من الأوضاع بمجموع «المُكافآت» لُكلِّ وضعٍ من أوضاع السِّلسلة. فإذا أعطيت الآلة غايةً ما وكانت الغاية مُعرَّفةً بدالَّة المنفعة أو المُكافأة، فإنها ستسعى لتنتهج سُلوكًا يُعزِّز من المنفعة المُتوقَّعة أو مجموع المكافآت المتوقَّع لها حسب مُتوسِّط النتائج المُحتملة وحساب أرجحيَّة حدوث كلٍّ منها. وهكذا يُعدُّ مجال الذكاء الاصطناعي الحديث جزئيًّا إعادة إحياءٍ لحُلم جون ماكارثي، مع استبدال المنافع والاحتمالات بالأهداف والمنطق.

في عام ١٨١٤، كتب عالم الرياضيات الفرنسي العظيم بيير سايمن لابلاس يقول: «نظرية الاحتمالات ما هي إلا اختزال البديهيات في معادلات التَّفاضل والتَّكامُل.»[57] وظلَّ الحال كما هو حتى ثمانينيات القرن الماضي حين طُوِّرت لُغة رسمية عملية وخوارزميات للتفكير لاستخدامها في عُلُوم الاحتمالات. وكانت تلك هي لغة «الشَّبكات البايزيَّة»[ج] التي قدَّمها جوديا بيرل. وعُمومًا، فالشَّبكات البايزيَّة ما هي إلا النظراء الاحتماليُّون لمنطق القضايا. وبالمثل، هناك أيضًا نظراء احتماليُّون للمنطق الإسنادي، بما في ذلك المنطق البايزي[58] وعدد هائل من «لُغات البرمجة الاحتمالية».

جاءت تسمية «الشَّبكات البايزيَّة» و«المنطق البايزي» تيمُّنًا بالقسِّ البريطاني المُبجَّل تُومس بايز الذي نُشرت مُساهمته الخالدة في الفكر الحديث عام ١٧٦٣ بعد وفاته بوقتٍ قصير على يد صديقه ريتشارد برايس،[59] والتي تُعرف الآن باسم «مُبرهنة بايز». تصف المُبرهنة، بصيغتها الحديثة التي قدَّمها بيير لابلاس، بطريقةٍ غاية في البساطة كيف يُمكن لاحتمالٍ «قبلي»؛ وهو الاعتقاد المبدئي الذي يتكوَّن لدى المرء فيما يتعلق بمجموعةٍ من الفرضيات المُحتملة، أن يُصبح احتمالًا «بعديًّا» كنتيجةٍ لُملاحظة بعض الأدلَّة. وكُلَّما توافدت أدلَّة جديدة، يُصبح الاحتمال البعدي احتمالًا قبليًّا جديدًا، وهكذا يستمر تحديثُ عملية مُبرهنة بايز إلى ما لا نهاية. وتُعدُّ هذه العملية أساسًا جوهريًّا، حتى إن الفكرة الحديثة عن العقلانيَّة التي ترى أنَّها وسيلة لتعزيز المنفعة المُتوقَّعة تُسمَّى أحيانًا بـ «العقلانيَّة البايزيَّة». وهي تفترض أنَّ الكيان العقلاني لدَيه معرفة بتوزيع للاحتمالات البعديَّة للأوضاع الحاضرة المُحتملة وللافتراضات المُستقبلية، استنادًا إلى كل خبراته السابقة.

كما طوَّر الباحثون في مجالات أبحاث العمليات ونظرية التَّحكُّم والذكاء الاصطناعي أنواعًا مُختلفةً من الخوارزميات لاتِّخاذ القرارات في ظلِّ وجود الارتياب وعدم اليقين، والتي

يعُود بعضها إلى خمسينيات القرن الماضي. وتلك الخوارزميات التي تُسمَّى خوارزميات «البرمجة الديناميكيَّة» هي النظراء الاحتماليُّون للبحث والتخطيط الاستباقي، ويُمكنها أن تُولِّد سُلُوكًا مثاليًّا أو قريبًا منه في جميع أنواع المشاكل العملية في المجال المالي وقطاع النَّقل والإمداد وغيرها من المجالات التي يلعب عدم اليقين فيها دورًا كبيرًا.[ج] توضع الغاية في تلك الآلات في صيغة دالة مكافأة، ويكون الناتج عبارةً عن «سياسة» تُحدِّد التَّصرُّف المُلائم لكل وضعٍ مُحتملٍ قد يتعرَّض له الكيان.

أما بالنسبة إلى المشاكل المُعقَّدة مثل لُعبة الطاولة ولُعبة جو حيث عدد الأوضاع هائل والمُكافأة لا تُمنح إلا بنهاية المُباراة، فهُنا البحث الاستباقي لن يُجدي نفعًا. وعوضًا عن ذلك، طوَّر باحثو الذكاء الاصطناعي أسلُوبًا يُسمَّى «التَّعلُّم المُعزَّز». وتتعلَّم خوارزميات التَّعلُّم المُعزَّز من الخبرات المُباشرة لإشارات المُكافأة في البيئة المُحيطة، تمامًا كما يتعلَّم الطفل الرضيع الوقوف على قدميه من المكافأة الإيجابية لكونه يقف مُعتدلًا ومن المكافأة السَّلبيَّة للوقوع أرضًا. وكما في خوارزميات البرمجة الديناميكيَّة، تكون الغاية الموضُوعة في خوارزمية التَّعلُّم المُعزَّز هي دالَّة المكافأة، ثمَّ تتعلَّم الخوارزمية قاعدة لتقدير قيم الأوضاع (أو أحيانًا قيم التَّصرُّفات). وهذه القاعدة يُمكن الجمع بينها وبين البحث الاستباقي قصير المدي نسبيًّا لتوليد سُلوكٍ ذي كفاءةٍ عالية.

كان أول نظام تعلُّمٍ مُعزَّز ناجح، هو برنامج آرثر صامويل للُعبة الدَّامة، الذي أثار ضجَّةً حين عُرض على شاشات التلفزيون عام ١٩٥٦. تعلَّم هذا البرنامج اللُّعبة من الصِّفر باللَّعب ضدَّ نفسه ثمَّ مُلاحظة مكافآت الفوز والخسارة.[60] وفي عام ١٩٩٢، طبَّق جيري تيزاورو الفكرة ذاتها وطوَّر برنامجًا للُعبة الطاولة، والذي حقَّق مُستوًى يُضاهي مستوى بطل العالم بعد ١٥٠٠٠٠٠ مُباراة.[61] وفي بدايات عام ٢٠١٦، استخدم برنامج «ألفا جو» وما تلاه من إصدارات من تطوير شركة ديب مايند أسلُوب التَّعلُّم المُعزَّز واللَّعب ضد النَّفس ليتمكَّن من هزيمة أمهر اللاعبين البشريين في ألعاب «جو» والشطرنج والشُّوجي.

تستطيع خوارزميات التَّعلُّم المُعزَّز أيضًا أن تتخيَّر التَّصرُّفات استنادًا إلى مُدخلاتٍ إدراكيَّة أوَّليَّة. على سبيل المثال، تعلَّم نظام «دي كيو إن» التابع لشركة ديب مايند كيفية لعب تسعة وأربعين لُعبة فيديو «أتاري» مختلفة من الصِّفر، بما في ذلك ألعاب «بونج» و«فري واي» و«سبيس إنفيديرز».[62] واستخدم فقط بكسلات الشاشة كمُدخلاتٍ، وعدد النقاط كإشارة مكافأة. وفي غالب الألعاب، تعلَّم نظام «دي كيو إن» أن يلعب أفضل من

أيِّ لاعبٍ بشري مُحترف رُغم أنَّه ليس لديه أيُّ سابق معرفةٍ بالزَّمان أو المكان أو العناصر أو الحركة أو السرعة أو الرِّماية. ومن الصَّعب أن نُحاول فهم ما الذي يفعلُه هذا النظام، إلى جانب الفوز في الألعاب.

إذا علمنا أنَّ رضيعًا في يومه الأول على الأرض قد تعلَّم كيفية لعب العشرات من ألعاب الفيديو بمُستوًى يفوق القدرات البشرية، أو صار بطل العالم في لُعبة جو أو الشطرنج أو الشُّوجي، فقد نظُنُّ أنَّ روحًا شيطانيةً تتلبَّسُه، أو أنَّه كائن ذو عقلٍ فضائي. لكن تذكَّر أنَّ كل تلك المهام هي أبسط بكثيرٍ من العالم الواقعي؛ فهي يُمكن ملاحظتها ملاحظةً كاملة ولها إطار زمني قصير، كما أنَّ لها فضاءات وضعية صغيرةً نسبيًّا وقواعد سهلة ويُمكن التَّنبؤ بها. وإذا أرخينا أيًّا من تلك الشروط، فهذا يعني أنَّ الطرق القياسية ستفشل.

على الجانب الآخر، الأبحاث الحالية تهدف على وجه التحديد إلى تخطِّي الطرق القياسية لكي تُصبح نظم الذكاء الاصطناعي قادرةً على العمل في فئات أكبر من البيئات. وإليكم مثالًا: في اليوم الذي كتبتُ فيه الفقرة السابقة، أعلنت شركة أوبن إيه آي أن فريقها المُكوَّن من خمسة برامج ذكاءٍ اصطناعي تعلَّم كيف يهزم فرقًا بشريةً مُحنَّكة في لُعبة «دُوتا ٢». (ولمثلي من غير المُطَّلعين، فلُعبة «دُوتا ٢» هي نُسخة مُطوَّرة من لُعبة «الدِّفاع عن آثار القُدماء»، وهي لُعبة استراتيجية آنيةُ الاستجابة من عائلة لُعبة «وور كرافت». وهي الآن أكثر الألعاب الإلكترونية تنافُسيَّةً وربحًا؛ فهي تُقدِّم جوائز بملايين الدولارات.) تتطلَّب لُعبة «دُوتا ٢» عملًا جماعيًّا وتواصُلًا بين اللاعبين، وزمانًا ومكانًا شبه مُتواصلين. فالمُباريات قد تصل إلى عشرات الآلاف من الفترات الزَّمنية، ويبدو أنه لا بُدَّ من وجود قدرٍ من التَّنظيم التَّسلسُلي للسُّلوك. وصف بيل جيتس هذا الإعلان بأنَّه «قفزة كبيرة للأمام في مجال الذكاء الاصطناعي».[63] وبعد عدة أشهرٍ، سحق إصدار مُحدَّث من البرنامج أمهر فريقٍ احترافي في العالم في لُعبة «دُوتا ٢».[64]

ألعاب مثل «جو» و«دُوتا ٢» تُعدُّ ساحة اختبارٍ مُمتازةٍ لأساليب التَّعلُّم المُعزَّز؛ لأنَّ دالة المكافأة تكون ضمن قواعد اللُّعبة. لكن العالم الواقعي أقل مُلاءمةً، وهناك العشرات من الحالات التي أدَّى التَّعريف الخاطئ للمكافآت إلى سلوكياتٍ غريبةٍ ومُفاجئة.[65] بعض هذه الحالات هي أخطاء بريئة مثل نظام مُحاكاة التَّطور والذي كان من المُفترض أن يُوجد كائنات سريعة الحركة، لكنَّ المطاف انتهى به وقد أنشأ كائناتٍ طويلةً كالنَّخل وتتحرَّك بسرعةٍ عن طريق السَّقوط مرارًا وتكرارًا.[66] وهناك حالات أخرى أقل براءةً مثل

أدوات تحسين مُعدَّل النَّقر على منصات التواصل الاجتماعي التي يبدو أنَّها تُضرم نار الفوضى في عالمنا.

آخر فئة من فئات برامج الكيان سأتحدَّث عنها هي أبسطها؛ وهي تلك التي تصل الإدراك مُباشرةً بالتَّصرُّفات دون أي تداولٍ أو تفكير وسيط. نُسمِّي مثل هذا النوع من البرامج في مجال الذكاء الاصطناعي بـ «برنامج الاستجابة اللاإرادية»، في إشارةٍ إلى ردود الفعل العصبية اللاإرادية البسيطة في البشر والحيوانات، والتي لا يتخلَّلُها أي تفكير.[67] ومثال ذلك هو استجابة «الرمش» التي تصلُ مُخرجات دوائر المُعالجة البسيطة في الجهاز البصري مُباشرةً بمنطقة العضلات التي تتحكم في حركة الجُفُون التي تكون على استعدادٍ إذا لاح طيف سريع مُقترب من العين في المجال البصري أن تُغمض بقوة. ويمكنك اختبار ذلك بنفسك الآن إذا حاولت (محاولةً بسيطةً) أن تُدخل إصبعك في إحدى عينيك. يُمكننا أن نُحوِّل نظام الاستجابة هذا إلى «قاعدةٍ» بسيطةٍ على النحو التالي:

إذا <لاح طيف سريع مُقترب في المجال البصري>، إذن <أغمض الجفنين>.

استجابة الرمش لا «تدري ما الذي تفعله»؛ فالغاية (حماية مُقلة العين من الأجسام الغريبة) لا تتجسَّد في أي مكان، وكذلك الحال بالنسبة للمعرفة (أيُّ طيفٍ سريعٍ يلُوح مُقتربًا يعني أنَّ جسمًا ما يقترب من العين، وأنَّ ذلك الجسم الذي يقترب من العين قد يضُرُّها). وهكذا فإنَّك عندما يُحاول الجانب الإرادي منك وضع قطرات في العين، فإنَّ الجزء اللاإرادي سيظلُّ يرمش ويُغمض الجفنين.

ومثال آخر لردِّ الفعل اللاإرادي هو «كبح الطَّوارئ»؛ عندما تتوقف السيارة التي أمامك فجأةً أو عندما يخطُو أحد المُشاة في الطريق. ليس من السَّهل أبدًا أن تُقرِّر بُسرعةٍ ما إذا كان استعمال المكابح ضرُوريًّا أم لا في تلك الحالة؛ في عام ٢٠١٨، عندما دهست سيارة اختبارية في وضع القيادة الذاتية أحد المُشاة، فسَّرت شركة أوبر الحادثة بأنَّ «إمكانية كبح الطوارئ لا تكون مُفعَّلةً حين تكون السيارة تحت تحكُّم الكمبيوتر؛ وذلك للتَّقليل من احتمال قيام المركبة بسلوك خاطئ.»[68] في تلك الحالة، كانت غاية المُصمِّم البشري واضحةً وضوح الشَّمس؛ وهي: «لا تقتُل المُشاة»، لكنَّ سياسة الكيان (لو كانت فُعِّلت) نفذتها بطريقةٍ خاطئة. ونؤكِّد هنا مرة أخرى أنَّ الغاية لم تُمثَّل في الكيان؛ فلا تُوجد سيارة ذاتيةُ القيادة في يومِنا هذا تفهم أنَّ الناس لا يُحبِّذُون أن يُدهسُوا.

للأفعال اللاإرادية دور أيضًا في العديد من المهام الأكثر اعتياديَّة مثل البقاء في إحدى حارات الطَّريق؛ فإذا ما حدث وحادت السيارة قليلًا عن الموضع المثالي لها في حارةٍ ما، فإنَّ نظام تحكُّمٍ بسيطًا يُمكنه أن يُحرِّك المِقود حركةً خفيفةً في الاتجاه المُعاكس لتصحيح المسار. ومقدار تلك الحركة سيعتمد على قدر انحراف السيارة. وهذا النَّوع من نظم التَّحكُّم إنما يُصمَّم عادةً لتقليل خطأ التَّتبُّع المُتراكم بمُرُور الوقت. إن المُصمِّم يستنتج قاعدةً للتَّحكُّم، آخذًا في الاعتبار افتراضاتٍ مُعيَّنة حول السُّرعة المُقرَّرة ومدى تقوُّس الطريق، والتي تعمل على تنفيذ عملية التقليل هذه تقريبيًّا.[69] وفي أجسادنا نظام مثل هذا يعمل طيلة الوقت وأنت واقف على قدميك، ولو حدث وتوقَّف ذلك النِّظام لخرَّ جسدك على الأرض في غضون ثوانٍ معدودة. وكما هو الحال بالنسبة إلى عملية الرمش، يكادُ يكون مستحيلًا أن تُوقف تلك الآلية عن العمل طواعيةً لينهار جسدُك على الأرض.

وهكذا فإن برامج الاستجابة اللاإرادية تُنفِّذ الغاية التي أودعها فيها المُصمِّم، لكنَّها في الوقت ذاته لا تعرف ماهيَّة تلك الغاية أو لماذا تتصرَّف على نحوٍ مُعيَّن. وهذا يعني أنَّها لا تستطيع اتِّخاذ القرارات بنفسها، بل يتَّخِذها شخص آخر نيابةً عنها ويُخطِّط لكل شيءٍ سلفًا؛ وهذا الشخص عادة ما يكون المُصمِّم البشري أو ربَّما يكون عملية التَّطور البيولوجي. من الصَّعب جدًّا أن تبني برنامجًا جيدًا من هذا النوع من خلال البرمجة اليدوية فقط، اللهم إلا لتنفيذ بعض المهام البسيطة جدًّا مثل لُعبة «إكس-أو» أو كبح الطَّوارئ. وحتى في تلك الحالات، يكون برنامج الاستجابة اللاإرادية جامدًا للغاية ولا يُمكنه تغيير سُلُوكه عندما تُشير الظُّروف إلى أنَّ السياسة التي طُبِّقت لم تعُد مُلائمة.

وإحدى الطَّرائق المُمكنة لبناء برنامج استجابة لا إرادية أكثر قوة هي عبر عملية للتَّعلُّم من الأمثلة.«د» وبدلًا من أن يُحدد المصمم البشري قاعدةً تُوضِّح للبرنامج كيف يتصرَّف أو يُزوِّده بهدفٍ أو دالة مُكافأة ما، فإنه يُمكنه أن يُغذِّيه بأمثلةٍ لمشاكل اتِّخاذ القرار، جنبًا إلى جنبٍ مع القرار الصحيح في كل مُشكلة. فمثلًا، يُمكننا أن نبني كيانًا يُترجِم من الفرنسية إلى الإنجليزية بتغذيته بأمثلةٍ لجُملٍ فرنسية جنبًا إلى جنبٍ مع الترجمة الإنجليزية الصحيحة. (لحُسن حظِّنا، يُصدر البرلمان الكندي وكذلك الخاص بالاتحاد الأوروبي الملايين من تلك الأمثلة سنويًّا.) بعد ذلك، تُعالج خوارزمية «تعلُّم مُوجَّه» الأمثلة لتُنتج قاعدةً مُعقَّدةً تأخذ أي جُملةٍ فرنسيةٍ كمدخلات فتُنتج ترجمةً لها بالإنجليزية. إن خوارزمية التَّعلُّم الرائدة حاليًّا في الترجمة الآلية هي ضرب من ضُرُوب

ما يُسمَّى بالتَّعلُّم المُتعمِّق الذي يُنتج قاعدةً على هيئة شبكةٍ عصبونيةٍ اصطناعية بمئات الطَّبقات وملايين المعاملات.«د» أما خوارزميات التَّعلُّم المُتعمِّق الأخرى، فقد تبيَّن أنها بارعة جدًّا في تصنيف العناصر في الصور والتَّعرُّف على الكلمات في إشارة كلامية. وهكذا فإنَّ الترجمة الآلية وتمييز العناصر المرئية والتَّعرُّف على الكلام تُعدُّ ثلاثةً من أهم المجالات الفرعية في مجال الذكاء الاصطناعي، وهي السَّبب وراء الحماس الزَّائد والنَّظرة المُتفائلة إلى مُستقبل التَّعلُّم المُتعمِّق.

يُمكنُني أن أجزم جزمًا شبه قاطعٍ بأنَّ التَّعلُّم المُتعمِّق هو ما سيقُودنا مُباشرةً إلى بناء ذكاء اصطناعي يُضاهي ذكاء الإنسان. ووجهة نظري في هذا الشأن، والتي سأشرحها لاحقًا، هي أنَّ التَّعلُّم المُتعمِّق ينقُصُه الكثير حتى يُحقِّق المطلوب،«د» لكن لنصُبَّ تركيزنا الآن على معرفة كيف تُستخدَم مثل هذه الطرائق في إطار النمُوذج القياسي للذكاء الاصطناعي حيث يُمكن للخوارزميات أن تُحسِّن من غايةٍ مُحدَّدة. بالنسبة إلى التَّعلُّم المُتعمِّق أو في واقع الأمر أي خوارزمية تعلُّمٍ مُوجَّه أخرى، فإن «الغاية التي جعلنا الآلة تسعى إلى تحقيقها» عادةً ما تكون تعظيم دقَّة التنبؤات أو بالطبع التَّقليل من الخطأ. وهذا القدْرُ يبدو واضحًا جليًّا، لكن يُمكن فهمُه بطريقتَين مختلفتَين طبقًا لدور القاعدة المُتعلمة في النِّظام برُمَّته. الدَّور الأول هو دور إدراكي محض؛ تُعالج الشَّبكة المُدخلات الحسِّية ثمَّ تمُدُّ بقيَّة النظام بالمعلومات في صورة تقديراتٍ احتماليةٍ لما تُدركه. فلو كانت الشبكة مثلًا هي خوارزمية تمييزٍ للعناصر المرئية، فلرُبَّما كانت المعلومات المُقدَّمة منها هي: «احتمالية ٧٠ بالمائة أنَّ العنصر هو كلب من سلالة «نورفولك تيرير»، و٣٠ بالمائة أنَّه كلب من سلالة «نورويتش تيرير».»[70] وعلى بقيَّة النظام أن يُقرِّر التَّصرف الخارجي استنادًا إلى تلك المعلومات. وهذا الهدف الإدراكي المحض لا يُمثِّل أدنى مُشكلةٍ إذا استوعبناه بالمعنى التالي: حتى النِّظام «الآمن» للذكاء الاصطناعي الخارق، في مقابل النِّظام «غير الآمن» المبني على أساس النمُوذج القياسي، يحتاج إلى وجود نظام إدراكٍ دقيقٍ ومعاير جيدًا قدر الإمكان.

تأتي المشكلة عندما ننتقل من الدَّور الإدراكي المحض إلى دور اتِّخاذ القرارات. فمثلًا، قد تُولِّد شبكة مُدرَّبة على تمييز العناصر المرئية تلقائيًّا تسمياتٍ للصور على موقعٍ من مواقع الإنترنت أو حسابٍ على منصَّةٍ من منصَّات التواصُل الاجتماعي. ولأنَّ نشر هذه التَّسميات يُعتبر تصرُّفًا له عواقبه، لذا تتطلَّبُ كُلُّ عملية توليدٍ للمُسمَّيات قرارًا تصنيفيًّا.

وما لم تكُن نتيجةُ كُلِّ قرارٍ من تلك القرارات مضمونةً تمامًا، حينها يكون لزامًا على المُصمِّم البشري أن يضع «دالة خسارة» تُوضِّح تكلفة الخطأ في تصنيف أحد العناصر من النَّوع «أ» على أنَّه عنصر من النَّوع «ب». وهذا يُفسِّر كيف واجهت شركة جوجل مُشكلةً عويصةً مع الغُوريلات. ففي عام ٢٠١٥، نشر مهندسُ برمجياتٍ يُسمَّى جاكي ألسنا تغريدةً على موقع «تويتر» يشتكي فيها أنَّ إمكانية تسمية الصُّور في خدمة «صور جوجل» قد وضعت تسميةً لصورةٍ له ولصديقه على أنَّهُما غوريلتان.[71] لا أحد يعرف كيف حدث مثل هذا الخطأ، لكن من شبه المُؤكَّد أنَّ خوارزمية جوجل لتعلم الآلة كانت مُصمَّمةً لتقليل دالَّة خسارةٍ مُحدَّدةٍ وثابتة. وفوق كلِّ هذا، فإنَّ تلك الدَّالة كانت تُعطي تكلفةً مُتساويةً لجميع الأخطاء. بعبارةٍ أخرى، إن تلك الدَّالة افترضت أنَّ تكلفة التَّصنيف الخطأ لشخصٍ على أنَّه غوريلا مثلُها كمثل تكلفة التَّصنيف الخطأ لكلبٍ من سُلالة «نورفولك تيرير» على أنَّه من سُلالة «نورويتش تيرير». ومن الجليِّ أنَّ تلك الدَّالة لم تكن دالة الخسارة الحقيقية لجوجل (أو مُستخدميها) كما تبيَّن من حجم الكارثة التي تسبَّبت بها على مُستوى العلاقات العامَّة.

وبما أنَّ هناك الآلاف من التَّسميات الممكنة للصُّور؛ فمن ثمَّ هناك الملايين من التَّكاليف الواضحة الخاصة بالتَّصنيف الخطأ لفئةٍ ما على أنَّها فئة أخرى. إن إيجاد كل تلك الأرقام وتحديدها مُقدَّمًا كان سيكون أمرًا غايةً في الصعُوبة على جوجل، حتى ولو حاولت فعله. لكنَّ الصَّواب هنا هو التَّسليم بعدم اليقين فيما يتعلَّق بالتكاليف الحقيقية للتَّصنيف الخطأ، والبدء بتصميم خوارزمية تعلُّمٍ وتصنيفٍ تكون ذات حساسيةٍ مُلائمةٍ للتَّكاليف وبعدم اليقين الذي يُحيط بها. مثل تلك الخوارزميات قد تسأل المُصمِّم البشريَّ في جوجل بين الفيْنة والفيْنة أسئلةً مثل: «أيهما أسوأ؛ التَّصنيف الخطأ لكلبٍ على أنَّه قطَّة، أم التَّصنيف الخطأ لإنسانٍ على أنَّه حيوان ما؟» بالإضافة إلى ذلك، قد ترفُض تلك الخوارزمية أن تضع أيَّ مُسمَّياتٍ لبعض الصُّور إذا وجدت أنَّ هناك قدرًا كبيرًا من الارتياب وعدم اليقين حول تكاليف التَّصنيف الخطأ.

في أوائل عام ٢٠١٨، تداول الناس أنَّ خدمة «صور جوجل» رفضت تصنيف صورةٍ لغوريلا. فرغم أنه أُعطي لها صورة غاية في الوضوح لغوريلا مع صغيرَين لها، فقد علَّقت قائلة: «امممم ... لم أستطع أن أُميِّز هذه الصورة جيدًا بعد.»[72]

أنا لا أُريد أن أُلَمِّح إلى أنَّ تبنِّي الذكاء الاصطناعي للنَّمُوذج القياسي كان اختيارًا سيئًا في ذلك الوقت؛ فقد بُذل قدر عظيم من العمل في تطوير العديد من النُّظم المنطقيَّة والاحتماليَّة والتعليميَّة المبنيَّة على ذلك النمُوذج. ونجد أنَّ العديد من النظم الناتجة هي فعلًا نُظُم مُفيدة؛ وكما سنرى في الفصل القادم، فما يزال هناك المزيد لنراه في المستقبل. على الجانب الآخر، فإنَّنا لا نستطيع أن نستمرَّ في اعتمادنا على أسلوب التَّجربة والخطأ لتحديد الأخطاء الجوهرية في دالة الغاية؛ فالآلات التي تزداد ذكاءً وتأثيرًا على مستوى العالم يومًا بعد يوم لن تسمح لنا بمثل هذه الرَّفاهية بعد الآن.

الفصل الثالث

كيف قد يتطوَّر الذكاء الاصطناعي في المُستقبل؟

(١) المُستقبل القريب

في الثالث من مايو عام ١٩٩٧، بدأت مُباراة شطرنج بين «ديب بلو»؛ الكمبيوتر الذي صمَّمته شركة آي بي إم ليلعب الشطرنج، وبين جاري كاسباروف؛ بطلُ العالم في الشطرنج ويُقال إنه أفضل لاعب شطرنجٍ بشري في التاريخ. وصفت حينها مجلَّة «نيوزويك» تلك المباراة بأنَّها «معركة الصُّمُود الأخيرة للعقل البشري». وفي الحادي عشر من مايو، بعد أن كانت النَّتيجة مُتعادلة ٢٫٥–٢٫٥، هزم «ديب بلو» كاسباروف في المُباراة النهائية. فهاجت وسائل الإعلام وماجت ووقفت الدنيا ولم تقعُد. وارتفعت القيمة السُّوقيَّة لشركة آي بي إم ١٨ مليار دولار بين عشيَّةٍ وضُحاها. وأُعلن على جميع الأصعدة أنَّ مجال الذكاء الاصطناعي قد أحرز تقدُّمًا هائلًا.

إذا نظرنا إلى الأمر من وجهة نظر أبحاث الذكاء الاصطناعي، فإن تلك المُباراة لم تُمثِّل أي طفرةٍ في المجال مُطلقًا. إن فوز «ديب بلو»، المُثير للإعجاب بلا شكٍّ، لم يكُن إلا استكمالًا لاتِّجاهٍ كان معرُوفًا مُنذ عُقُود. فالتَّصميم الأساسي لخوارزميات لعب الشطرنج وُضعت في عام ١٩٥٠ على يد كلود شانون،[1] ثمَّ طُوِّر تطويرًا كبيرًا في أوائل ستينيات القرن الماضي. بعد ذلك، ما فتئ تصنيف أفضل برامج لعب الشطرنج يتحسَّن باطِّرادٍ، على وجه الخصوص كنتيجةٍ لأجهزة الكمبيوتر الأكثر سرعة التي مكَّنت البرامج من تطوير أدائها والتطلع إلى مُستقبل أفضل. وفي عام ١٩٩٤،[2] كتبتُ أنا وبيتر نورفج التصنيف الرقمي لأفضل برامج الشطرنج من عام ١٩٦٥ فما بعده، وقد كان تصنيف جاري كاسباروف على ذلك المقياس هو ٢٨٠٥. بدأ ذلك التَّصنيف عند ١٤٠٠ في عام ١٩٦٥،

وأخذ يتطوَّر تطوُّرًا شبه مثاليٍّ على مدى ثلاثين عامًا. وبالاستنباط من الخطِّ البياني لما بعد عام ١٩٩٤، كانت التنبؤات تُشير إلى أنَّ الكمبيوتر سيكون قادرًا على هزم جاري كاسباروف عام ١٩٩٧، وهو ما حدث بالضبط.

أما بالنسبة إلى باحثي الذكاء الاصطناعي حينها، فإن الطَّفرات الحقيقية حدثت «قبل» أن يظهر «ديب بلو» على الساحة ويخطف الأضواء بثلاثين أو أربعين عامًا. وبالمثل، فإنَّ الشبكات الالتفافيَّة المُتعمِّقة ظهرت، وقد عولجت جميع عملياتها الرِّياضيَّة، قبل ما يزيد عن عشرين سنةً من بدء تصدُّرها للعناوين الرئيسية في وسائل الإعلام.

أما ما يتلَّقاه العامَّة من تصوُّراتٍ عن الطَّفرات في مجال الذكاء الاصطناعي عبر وسائل الإعلام (مثل فوز الآلات الساحق على البشر، وأنَّ إنسانًا آليًّا قد تجنَّس بجنسية المملكة العربية السعودية، وما إلى ذلك من أخبار)، فإنَّه لا ينطوي إلا على النَّذر اليسير من حقيقة ما يحدث في مُختبرات الأبحاث العالمية. فبداخل المُختبر، يقتضي البحث تفكيرًا مُطوَّلًا ونقاشاتٍ وكتابةً للصِّيغ الرياضيَّة على السُّبورات البيضاء. تُطرح الأفكارُ بلا انقطاعٍ؛ فمنها ما يُنحَّى جانبًا ومنها ما يُعاد اكتشافه. والفكرة الجيدة، التي قد تقودنا إلى طفرةٍ حقيقية، غالبًا ما تمُرُّ على أذهان الباحثين مرور الكرام في وقت طرحها، ثمَّ بعد ذلك قد تُفهم ويُنظر إليها أنَّها شكَّلت أساسًا لطفرةٍ جوهريةٍ في مجال الذكاء الاصطناعي؛ وذلك حين يُعيد اكتشافها شخصٌ آخر في وقتٍ أكثر ملاءمة. وحين تُجرَّب الأفكار، تُختبر مبدئيًّا لحلِّ مشاكل بسيطةٍ ليُنظر في أمر بديهياتها الأساسية أهي صحيحة أم لا، ثمَّ تُختبر لحلِّ مشاكل أصعب لنقف على حجم قُدراتها وإلى أي مدًى ستصل. وغالبًا ما تُخفق الفكرة المُفردة بنفسها في تقديم أي تحسُّنٍ ملحوظٍ في القُدرات، ولذا عليها أن تنتظر إلى أن تظهر فكرة أخرى فيُدمجا معًا ليُقدِّما قيمةً ما.

كل هذا العمل يكون مخفيًّا تمامًا عن الأنظار. ففي العالم الخارجي فيما وراء أبواب المُختبرات، يُصبح الذكاء الاصطناعي مرئيًّا فقط حين يتجاوز التَّراكُم التَّدريجي للأفكار وللبراهين الدَّالَّة على صحَّتها وفاعليتها عتبةً ما؛ أي النقطة التي يكون عندها من المفيد استثمار الأموال والجُهود الهندسية لبناء مُنتجٍ تجاري جديدٍ أو تقديم عرض مُبهر. حينها فقط، تُعلن وسائل الإعلام أنَّ طفرةً ما قد حدثت.

يمكن أن يتوقَّع المرء إذن أنَّ الأفكار الأخرى العديدة التي ما تزال جنينًا في رحم مُختبرات الأبحاث العالمية ستتخطَّى عتبة الرِّبحيَّة التِّجارية خلال السنوات القليلة المُقبلة. وسنرى هذا الأمر يكثُر حُدُوثه كُلَّما ازداد مُعدَّل الاستثمارات التِّجارية وزاد تقبُّل العالم

لتطبيقات الذكاء الاصطناعي أكثر. هذا الفصل سيُطلعُك على عينةٍ مما قد نراه واقعًا من طفراتٍ في هذا المجال في المُستقبل القريب.

وأثناء العرض، سأذكر بعض مساوئ تلك الطَّفرات التِّقنيَّة. وقد يجُول بذهنك العديد من مساوئها الأخرى، ولكن لا تحمل همًّا؛ فأنا سأُفرد الفصل القادم للحديث عن كل ذلك.

(١-١) بيئة الذكاء الاصطناعي

في البداية، كانت البيئة التي عملت بها معظم أجهزة الكمبيوتر فراغًا لا شكل له؛ فمُدخلاتها الوحيدة كانت تأتي من البطاقات المثقُوبة وكانت الطريقة الوحيدة لإنتاج المخرجات هي طباعة الرموز عبر طابعة سطرية. رُبما لهذا السبب، كان يرى مُعظمُ الباحثين الآلات الذكية على أنَّها آلات تجيب على الأسئلة، ولمْ ينتشر المفهوم الحالي للآلات بأنَّها «كيانات ذكية» تُدرك وتتصرَّف في بيئةٍ ما إلا في ثمانينيات القرن الماضي.

عندما اختُرعت شبكة الويب العالمية في تسعينيات القرن الماضي، فتحت بابًا واسعًا لعالمٍ جديدٍ أمام الآلات الذكية لتتحرك فيه. واستُحدثت كلمة جديدة وهي «سوفت بوت» لوصف «روبوتات» البرمجيات التي تعمل بالكامل في بيئةٍ برمجيَّةٍ مثل شبكة الويب. هؤلاء الآليُّون يطلعون على صفحات الويب ثمَّ ينتجون استجابةً عن طريق إنتاج مجموعاتٍ من الرموز وعناوين الصفحات وغيرها من الأشياء.

ازداد عدد شركات الذكاء الاصطناعي ازديادًا كبيرًا خلال الفترة التي حدث فيها ما يُسمى بـ «فقاعة الإنترنت» والتي كانت ما بين عامي ١٩٩٧ و٢٠٠٠، مما وفَّر القُدرات الأساسية للبحث والتجارة الإلكترونية، بما في ذلك تحليل الرَّوابط ونظم التوصية ونظم بناء السُّمعة والتسوق القائم على مُقارنة السِّلع وتصنيف المُنتجات.

وفي بداية الألفيَّة الجديدة، مهَّد الانتشار الواسع للهواتف الجوَّالة، بما فيها من ميكروفونات وكاميرات ومقاييس تَسارُعٍ ونظُم تحديد مواقع، الطريق أمام نُظم الذكاء الاصطناعي لتتغلغل في حياة البشر اليومية، ثمَّ ها نحن نرى «السماعات الذكية» مثل «إيكو» التابعة لشركة أمازون و«هوم» التابعة لشركة جوجل و«هوم بود» التابعة لشركة أبل وقد أكملت هذه العملية.

وبحُلول عام ٢٠٠٨ تقريبًا، تخطَّى عدد الأشياء المُتَّصلة بالإنترنت عدد البشر المُتَّصلين بها، في نقلةٍ يُشير إليها البعض بأنَّها كانت بداية مفهوم «إنترنت الأشياء». وتلك الأشياء

تتضمن السيارات والأجهزة المنزلية وإشارات المرور وآلات البيع والترموستات والطَّوافات الرباعيَّة والكاميرات والحساسات البيئية والروبوتات، وجميع أنواع السِّلع المادِّية في كلٍّ من عملية التَّصنيع ونظام التَّوزيع والبيع بالتَّجزئة. إن هذا يتيح لنظم الذكاء الاصطناعي وصولًا أكبر بكثيرٍ إلى العالم الواقعي.

وأخيرًا، التطوُّرات التي حدثت في الإدراك مكَّنت الروبوتات المدعومة بنظم ذكاء اصطناعي من مُغادرة المصانع حيث كانت تعتمد على ترتيباتٍ ثابتةٍ ومُقيِّدةٍ للأشياء، وباتت الآن في قلب العالم الواقعي المليء بالفوضى والمُفتقر للنُّظم حيث تطَّلع كاميراتها على أشياء شيِّقةٍ وأكثر إمتاعًا.

(١-٢) السَّيارات الذاتية القيادة

في أواخر خمسينيات القرن الماضي، تصوَّر جون ماكارثي أنَّ مركبة مؤتمتة قد تُقلُّه إلى المطار في يومٍ من الأيام. وفي عام ١٩٨٧، أجرى إرنست ديكمانز تجربةً لشاحنةٍ ذاتية القيادة من إنتاج شركة مرسيدس على شبكة الطرق السَّريعة الألمانية «أوتوبان»، وقد كانت تلك الشاحنة قادرةً على الالتزام بالسَّير في إحدى حارات الطريق، والسَّير خلف سيارةٍ أخرى، وتغيير الحارات وتخطِّي السيارات التي أمامها.[3] بعد تلك التجربة بأكثر من ثلاثين عامًا، لا يُوجد بعدُ سيارات ذاتية القيادة باستقلاليةٍ كاملة، ولكنَّنا أوشكنا على تحقيق ذلك. والجدير بالذِّكر أنَّ التَّركيز على التَّطوير قد انتقل منذ مدةٍ طويلة من مختبرات الأبحاث الأكاديميَّة إلى الشَّركات الكبيرة. وبحلول عام ٢٠١٩، أتمَّت أفضل السيارات الاختبارية الذاتية القيادة ملايين الأميال من القيادة على الطُّرُق العامة (وملياراتٍ من الأميال في نظم مُحاكاة القيادة) دون أي حوادث خطيرة.[4] ولكن للأسف، هناك مركبات أخرى ذاتية القيادة أو شبه ذاتية القيادة قتلت العديد من الأفراد.[5]

والسُّؤال هُنا: لم استغرقنا كل ذلك الوقت لتحقيق القيادة الذَّاتية الآمنة؟ السَّبب الأول هو أنَّ مُتطلَّبات الأداء كثيرة وتستلزم الحرص والدِّقَّة. مثلًا في الولايات المتحدة، يتكبَّد السائق البشري تقريبًا حادثةً واحدةً مُميتةً في كُلِّ مائة مليون ميلٍ يقطعُها بسيارته. هذا بدوره يضعُ معيارًا عاليًا لقبُول المركبات الذاتية القيادة التي عليها إذن أن تُحقِّق مُستوًى أعلى من ذلك؛ رُبَّما بمُعدَّل حادثةٍ مُميتةٍ في كل مليار ميلٍ أو خمسة وعشرين عامًا من القيادة بمُعدَّل أربعين ساعةٍ أسبوعيًّا. أما السبب الثاني فهو ببساطةٍ فشلُ وسيلة

التَّحايُل المُتوقَّعة، المتمثلة في إسناد التحكم في السيارة للسائق البشري حين تكون المركبة مُشوَّشةً ولا تقدر على اتِّخاذ القرار أو خارج ظروف العمل الآمنة التي صُمِّمت للعمل فيها. فعندما تقود السيارة نفسها، سرعان ما ينفصل البشر عن ظروف القيادة المُباشرة ولا يُمكنُهم استعادة وعيهم بالبيئة المُحيطة استعادةً سريعةً تكفي لتولِّي القيادة بأمان. زد على ذلك أن الرُّكاب غير القادرين على القيادة في السيارة ورُكاب سيارات الأجرة في المقعد الخلفي يكونون في وضعٍ لا يسمح لهم بتولِّي القيادة إن حدث خطأ ما.

تسعى المشاريع الحالية للوصول إلى المُستوى الرابع من مُستويات القيادة الذاتية التي وضعتها جمعية مهندسي المركبات،[6] والذي يعني أنَّ المركبة يجب أن تكون قادرةً في جميع الأوقات على القيادة الذاتية أو التَّوقُّف الآمن أخذًا في الاعتبار القُيُود الجغرافية وحالات الطَّقس. ولأنَّ حالتَي الطَّقس والمُرُور يمكن أن يتغيَّرا، كما يُمكن أن تنشب ظُرُوف استثنائيَّة لا يُمكن لمركبةٍ من المستوى الرابع التَّعامُل معها، لهذا يجب أن يُوجَد سائق بشري في السيارة، وأن يكون مُستعدًّا لتولِّي القيادة إن لزم الأمر. (المستوى الخامس — القيادة الذاتية الكاملة — لا يتطلَّب وجود أي سائقٍ بشري مطلقًا، لكن هذا المُستوى هو أصعب في الوصول إليه مما قبله.) المستوى الرابع من القيادة الذاتية يتجاوز المهام البسيطة مثل اتِّباع الخطوط البيضاء وتفادي العقبات. فالمركبات عليها أن تُقيِّم النَّوايا والمسارات المُستقبليَّة المُحتملة لجميع الأشياء على الطريق، بما في ذلك الأشياء التي تقع خارج نطاق الرؤية؛ وذلك اعتمادًا على الملاحظات الحالية والماضية. ثمَّ عليها أن تستخدم البحث الاستباقي لتجد مسارًا يُحقِّق على النحو الأمثل مزيجًا من الأمان والتَّقدم نحو الهدف. بعض المشاريع تُجرِّب الآن مناهج مُباشرةً أكثر استنادًا إلى التَّعلُّم المُعزَّز (يتمُّ ذلك في نظم المُحاكاة طبعًا) والتَّعلُّم المُوجَّه من تسجيلاتٍ لمئات السائقين البشريين، ولكن تلك المناهج يبدو أنَّها من غير المُحتمل أن تُحقِّق المستوى المطلوب من الأمان.

المنافع المُحتملة للمركبات ذات القيادة الذاتية الكاملة هائلة. سنويًّا، يموت قرابة ١٫٢ مليون شخصٍ في حوادث السيارات حول العالم ويُعاني عشرات الملايين من إصاباتٍ خطيرة بسببها. وأحد الأهداف المعقولة لسيارات القيادة الذاتية هو تقليل تلك الأرقام إلى العُشر. كما تتنبَّأ بعض التحليلات بانخفاضٍ كبيرٍ في تكلفة المواصلات، وهياكل مواقف الانتظار والازدحامات المرورية، ومعدَّل التلوث. سيتحوَّل سكان المدن عن السيارات الشخصية والحافلات الكبيرة إلى تشارك المركبات الكهربية الذاتية القيادة المُنتشرة في كل مكانٍ والتي تُقدِّم خدمة توصيلٍ من الباب إلى الباب عبر شبكاتٍ نقلٍ عامٍّ عالية السرعة

بين المحطات الرئيسية.[7] تلك التكلفة المُنخفضة التي تقدر بثلاثة سنتات لكل ميلٍ يقطعه المسافر، ستعمل معظم المدن لتوفير تلك الخدمة مجانًا – بينما تفرض على الركاب وابلًا من الإعلانات التي لا تنقطع طوال الرحلة.

بلا شك، إذا أردنا أن نجني كل تلك المزايا، على أرباب الصناعة أن ينتبهوا جيدًا للمخاطر المُحتملة. إذا ارتفع عدد ضحايا المركبات الاختبارية السيئة التصميم، قد تُوقف الجهات التنظيمية خُطط الانتشار المُعدَّة أو ربما يفرضون معايير غايةً في الصَّرامة قد لا تُحقَّق إلا بعد عقودٍ كثيرة.[8] وبالطبع، رُبَّما يُقرِّر الناس ألا يشتروا أو يركبوا المركبات الذاتية القيادة إلا إذا ثبت أمانها. أظهر استفتاء أُجري عام ٢٠١٨ انخفاضًا حادًّا في مُستوى ثقة المُستهلكين في تقنية المركبات الذاتية القيادة؛ وذلك بالمقارنة باستفتاءٍ آخر تمَّ في عام ٢٠١٦.[9] وحتى إن كانت التقنية ذاتها ناجحةً، فإنَّ التحوُّل إلى مرحلة القيادة الذاتية الواسعة النطاق ستكون مرحلةً صعبةً؛ فمهارات القيادة البشرية قد تضعف أو تختفي، وقد تختفي بالكُلِّية قيادة الفرد المتهورةً والضارةً بالمجتمع لسيارته بنفسه.

(١-٣) المُساعد الشَّخصي الذَّكي

معظم قُراء هذا الكتاب من المفترض أن يكونوا قد جرَّبوا المُساعد الشَّخصي غير الذكي: السماعة الذكية التي تُنفِّذ أمرًا بشراء شيءٍ سمعه من شخصٍ ما على التلفزيون، أو نظام الدردشة على الهاتف الجوال الذي يُجيب على شخصٍ كتب: «استدع لي سيارة إسعاف!» بالآتي: «حسنًا! سأدعوك من الآن «آن سيارة إسعاف»». مثل تلك النظم هي في الأساس عبارة عن واجهاتٍ صوتيةٍ للتطبيقات ومحرِّكات البحث؛ وهي مبنية في العُمُوم على قوالب الردود الجاهزة، وهو أسلوب قديم يعود إلى نظام «إليزا» الذي ظهر في مُنتصف ستينيات القرن الماضي.[10]

تلك النظم البدائية لها عيوب تندرج تحت ثلاثة أقسام: الدِّراية والمحتوى والسِّياق. «عيوب الدِّراية» تعني أنَّها تفتقد الوعي الحسي بما يحدث حولها؛ فمثلًا، قد يُمكن لتلك النظم أن تسمع ما يقوله المُستخدِم، لكن لا يُمكنها أن ترى لمن يُوجِّه المُستخدمُ حديثه. و«عيوب المحتوى» تعني أنَّها ببساطةٍ لا تستطيع فهم معنًى ما يقوله المُستخدِم أو يكتبُه، حتى ولو لديها دراية به. أما «عيوب السِّياق» فتعني أنَّها لا تملكُ القدرة على مُتابعة أي سياقٍ والتَّفكُّر بما يحتويه من أهدافٍ وأنشطةٍ وعلاقاتٍ تُشكِّل أجزاء الحياة اليومية.

رغم تلك العيوب، فإنَّ السماعات الذكية ومُساعدات الهواتف الجوَّالة تُقدِّم ما يكفي من قيمةٍ للمُستخدم ليُدخلها مئات الملايين من الناس بُيوتهم ويحملوها معهم في جُيوبهم. هذه النُّظُم يمكن النظر إليها على أنَّها أحصنة طروادة لمجال الذكاء الاصطناعي. ونظرًا لأنها مدمجة في نسيج حياة الكثير من الناس، فإن كلَّ تطورٍ في قدراتها، مهما كان صغيرًا، يُساوي المليارات من الدولارات.

وكما هو معروف؛ فالتطويرات تأتي بسرعةٍ وكثافة. وربَّما كان أهمها هو القدرة الأساسية على فهم المُحتوى؛ أي مثلًا فهم أنَّ جُملة «جون في المستشفى» لا يُردُّ عليها فقط بجملة «أرجو أن يكون بخير!» لكنها جُملة تحوي معلومةً حقيقةً وهي أن ابن المُستخدم ذا الثماني سنواتٍ في مُستشفًى قريبٍ وربما كان مُصابًا أو مريضًا وحالته خطيرة. تُعدُّ قدرة نظم الذكاء الاصطناعي على الوصول إلى البريد الإلكتروني والرسائل النَّصيَّة بالإضافة إلى المكالمات الصوتية والمحادثات المنزلية (من خلال السماعة الذكية) عاملًا مهمًّا في الحصول على ما يكفي من المعلومات لبناء تصوُّرٍ كاملٍ نسبيًّا عن حياة المُستخدم؛ تلك المصادر رُبما تُعطي معلوماتٍ أكثر مما كان متاحًا لكبير الخدَم الذي يعمل لدى أسرةٍ من طبقة النُّبلاء في القرن التاسع عشر، أو المساعد التنفيذي الذي يعمل برفقة رئيس إحدى الشركات المعاصرة.

تلك المعلومات الأساسية ليست كافية بلا شك. ولتكون تلك المعلومات مُفيدة بحق، على المُساعد الذكي أن يكون على علمٍ ببديهيات العالم وكيف يعمل: فالطفل الذي في المُستشفى لا يكون في المنزل في الوقت ذاته؛ والرعاية الطبية لذراعٍ مكسورة نادرًا ما تدوم لأكثر من يومٍ أو يومين؛ وأنَّ مدرسة الطفل يجب أن تعرف بذلك الغياب المُتوقَّع؛ وغيرها من البديهيات. تلك المعرفة تُمكِّن المُساعد الذكي من أن يُتابع سياق الأشياء التي لا يراها مُباشرةً؛ وهي مهارة أساسية للنظم الذكية.

أعتقد أنَّ ما أشرتُ إليه من قدراتٍ في الفقرة السابقة هي قدرات قابلة للتنفيذ في ظلِّ التقنية الحالية للتفكير الاحتمالي،«ج» لكن هذا سيتطلَّب جهدًا جبارًا لإنشاء نماذج لجميع أنواع الأحداث والتعاملات التي تُشكِّل حياتنا اليومية. حتى الآن، هذه الأنواع من مشاريع إنشاء نماذج البديهيات بوجهٍ عامٍّ لم تُدشَّن بعد (اللهمَّ إلا ربما في بعض النظم السِّرية للتحليلات الاستخباراتية والتخطيط العسكري)، وهذا راجع إلى تكلفتها الباهظة ونتائجها غير المُؤكَّدة. أما الآن، فمشاريع مثل هذه في استطاعتها الوصول إلى مئات الملايين من المُستخدِمين، وبذلك تقلُّ مخاطر الاستثمار وتزيد فُرص المكاسب المُحتملة على نحوٍ أكبر.

أضف على ذلك أنَّ إمكانية الوصول إلى عددٍ كبيرٍ من المُستخدمين يزيد من سُرعة تعلُّم المُساعد الذكي واكتسابه لكلِّ ما ينقُصُه من معرفة.

على هذا النَّحو، نستطيع أن نترقَّب رؤية مُساعداتٍ ذكيةٍ تقدِر، لقاء حفنةٍ من البنسات كل شهر، أن تُساعد المُستخدِمين في إدارة يومهم بما فيه من أنشطةٍ كثيرةٍ ومتنوِّعة مثل: المواعيد والرحلات، والتَّسوق للحصول على احتياجات المنزل ودفع الفواتير ومُساعدة الأطفال في الفُرُوض المدرسية، وفرز رسائل البريد الإلكتروني والمكالمات الواردة، والتنبيهات، وإعداد الوجبات، ورُبما نشطح بأحلامنا ونأمُل أن تُساعدهم أيضًا في إيجاد مفاتيحهم الضائعة. كل تلك المهارات لن تكون مُبعثرةً بين العديد من التطبيقات؛ بل ستكون جميعها تحت مظلَّة كيانٍ واحد ومُتكامل يُمكنه أن يستفيد من جميع فُرص التَّضافُر والتآزُر المُتاحة، فيما يُسمِّيها العسكريون بـ «الصورة العملياتية العامة».

يتضمَّن القالب التَّصميمي العام لأي مُساعدٍ ذكي المعرفة المُسبقة للأنشطة البشرية والقدرة على استخلاص المعلومات من بين تدفقات البيانات الإدراكية والنَّصيَّة. ويتضمَّن أيضًا عملية تعلُّمٍ لتهيئة المُساعد لظُرُوف المُستخدم الخاصة. وهذا القالب العام يُمكن تطبيقه في ثلاثة مجالاتٍ كبرى أخرى على الأقل: الصِّحة والتَّعليم والشُّئون المالية الشخصية. وفي تلك التطبيقات، على النِّظام أن يُتابع جسد المُستخدم أو عقله أو حسابه المصرفي (إذا ما فسَّرنا مهامه تفسيرًا واسعًا). وكما هو الحال مع المُساعدات الخاصة بالحياة اليومية، فإن التَّكلفة الأولية لبناء المعرفة العامة الضَّرورية لكل مجالٍ من تلك المجالات الثلاثة ستُسدَّد تدريجيًّا من خلال الوصول إلى مليارات من المُستخدِمين.

في المجال الطبي، مثلًا، نحن البشر لنا جميعًا نفس وظائف الأعضاء إلى حدٍّ كبير، والمعرفة المُفصَّلة لكيفية عمل تلك الوظائف قد شُفرت بالفعل بشكل تفهمه الآلات.[11] فالنظم إذن ستتكيَّف مع خصائصك الفردية ونمط حياتك الشخصي لتُقدِّم لك اقتراحاتٍ وقائيةً وتحذيراتٍ مُبكِّرةً للأمراض والمشاكل.

وفي المجال التعليمي، فقد كانت هناك وعود بداية من الستينيات من القرن الماضي ببناء نظم تدريس ذكية،[12] لكنَّ التَّقدُّم الحقيقي في تلك الفكرة غاب طويلًا وطال انتظاره. والأسباب الرئيسية لذلك التأخر هي عيوب المحتوى وعيوب الوصول؛ فغالبية نظم التَّدريس لا تفهم المحتوى الذي يُراد منها شرحه، ولا تستطيع أن تنخرط في تواصُل ثنائي مع الطلاب لا بالكلام ولا بالكتابة. (أتخيَّل نفسي وأنا أُدرِّس نظرية الأوتار التي لا

أفهمها باللغة اللاوية التي لا أتحدثها.) لكنَّ التَّقدم الحديث في تقنية التَّعرف على الكلام يعني أنَّ المدرسين الآليين يمكنهم أخيرًا التَّواصل مع طلابهم في سنوات تعليمهم الأولى. أضف على ذلك أنَّ تقنية التفكير الاحتمالي يمكنها الآن أن تُتابع ما يعرفه الطلاب وما لا يعرفونه؛[13] ومن ثمَّ يمكن أن تُحسِّن من طرائق التَّدريس لتحقيق أكبر تحصيلٍ للمعرفة. وفي هذا الشأن تُقدِّم مسابقة «إكس برايز» العالمية للتعلم التي بدأت عام ٢٠١٤، جائزةً قدرها ١٥ مليون دولار لمن يُصمِّم «برمجيات مفتوحةَ المصدر ذات قابليةٍ للتَّطوير تُمكِّن الأطفال في الدول النامية من تعليم أنفسهم مبادئ القراءة والكتابة والحساب في خلال ١٥ شهرًا». والنتائج التي قدَّمها الفائزان في المسابقة وهما «كيت كيت سكول» و«وان بليون»، تُشير إلى أنَّ ذلك الهدف المرجُو قد حُقِّق على نحوٍ كبير.

أما في مجال الشُّئون المالية الشخصية، فسوف تُتابع النظم سير الاستثمارات، وتدفُّق مصادر الدَّخل، والنَّفقات الإلزامية والاختيارية، والديون وسداد الفوائد، ومُدَّخرات الطوارئ وما إلى ذلك، تمامًا كما يُتابع المُحلِّلُون الماليون الشئون المالية للشركات وفُرصها المُستقبليَّة. ولو تكامل هذا مع عمل الكيان الذكي الذي يُدير شئون الحياة اليومية، فسيُوفِّر ذلك فهمًا أعمق وأدق للجوانب المالية، ورُبما كان من فوائد ذلك أن يحرص النِّظام على خصم أي عقوباتٍ وقعت على الأطفال الأشقياء من مصروفهم الشهري قبل إعطائهم إيَّاه. وهكذا، يمكن للمرء أن يتوقع الحصول على نوعية النصائح المالية اليومية التي كانت مقتصرة في السابق على الأشخاص ذوي الثراء الشديد.

إذا لم تفزع وأنت تقرأ تلك الفقرات السابقة وتُثار بداخلك تساؤلات عديدة حول خصوصيتك، فغالبًا أنت لا تُتابع الأخبار يا صديقي! ورغم ذلك، هناك عدة فُصُول في حكاية الخصوصية هذه. أولًا: دعنا نتساءل: هل يُمكن للمُساعد الشخصي أن يكون مُفيدًا حقًّا إذا لم يعرف عنك أي شيء؟ غالبًا الإجابة هي لا. ثانيًا: هل يُمكن للمُساعد الشخصي أن يكون مفيدًا حقًّا إذا لم يستطع أن يجمع المعلومات من المُستخدِمين ليتعلَّم منها وتزيد معرفته بالبشر عمومًا، والبشر الذين يُشبهُونك؟ الإجابة في الأغلب هي لا أيضًا. ولكن هل تعني هاتان النقطتان أنَّ علينا أن نتخلَّى عن خصوصيتنا إذا ما أردنا أن نستفيد من الذكاء الاصطناعي في حياتنا اليومية؟ سأُجيب هنا أيضًا بلا. والسبب وراء ذلك هو أنَّ خوارزميات التَّعلُّم يُمكن أن تُعالج بياناتٍ «مُشفَّرة» باستخدام أساليب الحوسبة الآمنة المُتعدِّدة الأطراف، بحيث يُمكن أن يستفيد المُستخدمون من تجميع البيانات دون

التَّفريط في خصوصيتهم مُطلقًا.[14] وهل سيُطبِّق مُصمِّمو البرمجيات تقنيات الحفاظ على الخصوصية طوعًا من أنفسهم، دون إلزام قانوني؟ هذا ما سيتَّضح في المُستقبل. لكن ما يبدو أمرًا حتميًّا ولا مفرَّ منه هو أنَّ المُستخدِمين سيثقون في المساعد الشَّخصي الذكي فقط حين يكون ولاؤه الأساسي للمُستخدم أولًا، لا للشَّركة التي صمَّمته.

(٤-١) المنازل الذكيَّة والروبوتات المنزلية

عُرض ونوقش مفهوم المنازل الذكيَّة قبل عدة عقودٍ. في عام ١٩٦٦، بدأ جيمس سذرلاند؛ المهندس في شركة «وستينج هاوس»، تجميع ما تبقى من أجزاء كمبيوتر سابق طورته شركته لبناء «إيكو» التي تعدُّ أول وحدة تحكُّمٍ خاصَّة بالمنازل الذكيَّة.[15] ولكن للأسف، كانت «إيكو» تزن ثمانمائة باوند، وتستهلك ٣,٥ كيلووات من الكهرباء وكانت تتحكَّم فقط في ثلاث ساعاتٍ رقمية وهوائيِّ التلفزيون. أما النُّظُم اللاحقة، فقد كانت تطلُب من المُستخدِمين أن يتعاملوا مع واجهات تحكُّمٍ شديدة التَّعقيد. وكما يُمكنُك أن تتوقَّع، لم تنجح قط.

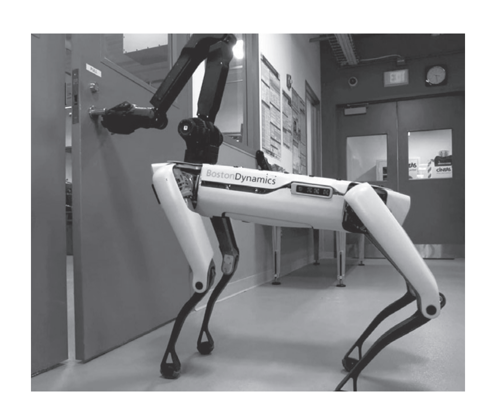
وبداية من تسعينيات القرن الماضي، ظهرت عدَّة مشاريع طموحة حاولت أن تُصمِّم منازل تُدير نفسها بنفسها مع تدخُّلٍ بشريٍّ بسيط؛ وذلك باستخدام تقنيات تعلُّم الآلة للتأقلُم مع نمط حياة ساكنيها. ولتكون تلك التَّجارب ذات معنًى، فقد اضطُرَّ بعض الأشخاص إلى أن يعيشوا في تلك البيوت. ومع الأسف، إن عدد القرارات الخاطئة التي اتَّخذتها تلك النظُم قد جعل منها نُظمًا عديمة الفائدة، بل أسوأ من ذلك بكثير؛ فقد انخفضت جودة حياة سُكانها بدلًا من أن تزيد. فمثلًا، اضطُرَّ سُكان منازل مشروع «ماف هوم»[16] عام ٢٠٠٣ بجامعة ولاية واشنطن أن يمكُثُوا في الظَّلام أغلب الوقت إذا كان في البيت زُوَّار وظلُّوا برفقتهم بعد وقت النوم المُعتاد.[17] ومثل ما حدث مع المُساعد الشَّخصي غير الذَّكي، مثل هذه الإخفاقات هي نتيجة نقصٍ في إدراك أنشطة السُّكان وعدم القدرة على فهم ومُتابعة ما يحدث في المنزل.

البيت الذَّكي الحقيقي، المُجهَّز بالكاميرات والميكروفونات والمتمتِّع بالقدر الأساسي من الإدراك وقدرات التَّفكير، يُمكنه فهم ما يفعله سُكانه؛ ألديهم زُوَّار؟ أهُم يأكلون أم هُم نيام؟ أيُشاهدون التِّلفزيون أم يقرءون؟ أيُمارسون التَّدريبات أم يُجهِّزون لرحلةٍ طويلة أم مُلقَون على الأرض دون حراك بعد أن سقطوا؟ وبالتَّنسيق مع المُساعد الشَّخصي

الذَّكي، يُمكن للمنزل أن يُكوِّن تصوُّرًا جيدًا للغاية عمَّن سيكون في المنزل ومن سيكون خارجه وفي أي وقتٍ تحديدًا، ومن يتناول الطَّعام وأين بالضَّبط، وما إلى ذلك من أمور. وهذا الفهم سيسمح له بالتَّحكُّم في تدفئة المنزل وإضاءته وستائر نوافذه وأنظمته الأمنية، وسيتيح له أن يُرسل التَّنبيهات التذكيرية في مواعيدها الصحيحة، وأن ينبه السُّكان أو يتَّصل بخدمة الطَّوارئ إذا ما حدثت مُشكلة. جدير بالذكر أن بعض المُجمَّعات السَّكنيَّة المبنية حديثًا في الولايات المُتَّحدة واليابان تُطبِّق بالفعل مثل هذه التقنيات.[18]

إن القيمة المتحقِّقة من المنازل الذَّكيَّة محدودة بسبب أنظمتها المُشغِّلة؛ فالنظم الأبسط تركيبًا (مثل الثرموستات المُؤقَّتة، والأضواء الحساسة للحركة، وأجهزة الإنذار الخاصة بالسَّرقة) يُمكنُها تنفيذ مُعظم المهام بطرائق قد تكون أكثر توقُّعًا وإن كانت أقل حساسية للسِّياق. والمنزل الذَّكي لا يُمكنه أيضًا طيُّ الغسيل أو تنظيف الأطباق أو التقاط الجريدة من أمام باب المنزل؛ فمثل هذه المهام تتطلَّب إنسانًا آليًّا في هيئةٍ ماديةٍ ليُقدم عليها.





شكل ٣-١: على اليمين الروبوت «بريت» يطوي المناشف، وعلى اليسار الروبوت «سبوت ميني» من شركة بوسطن ديناميكس يفتح الباب.

من المُحتمل ألا ننتظر طويلًا؛ فقد أظهرت الروبوتات عددًا كبيرًا من المهارات المطلوبة. مثلًا: في مُختبر صديقي بيتر أبيل بمختبر بيركلي لتعليم الروبوتات، أصبح

الروبوت «بريت» (وهو الروبوت الذي أعدَّه مُختبر بيركلي لتولِّي المهام البسيطة والمُملَّة) يطوي المناشف ويرصُّ بعضها فوق بعضٍ منذ ٢٠١١، بينما يستطيع الروبوت «سبوت ميني» من شركة بوسطن ديناميكس صُعود السلالم وفتح الأبواب (انظر الشكل ٣-١). كما أنَّ العديد من الشركات تبني حاليًّا روبوتات لطهو الطعام، رُغم أنَّها تتطلب تجهيزات خاصةً ومكوناتٍ جاهزةً، ولن تعمل في المطابخ العادية.[19]

ومن بين المهارات المادِّية الأساسية الثلاث المطلوبة للروبوت المنزلي حتى يكون مفيدًا — الإدراك وسهولة الحركة والبراعة — تُعدُّ المهارة الأخيرة أكثرها إشكالًا. والأمر كما عبَّرت عنه ستيفاني تاليكس أستاذة علم الروبوتات بجامعة براون فقالت: «مُعظم الروبوتات لا تستطيع التقاط مُعظم الأشياء في مُعظم الأوقات.» جزء من هذا يرجع إلى مشكلةٍ في حاسة اللمس، وجزء آخر يرجع إلى مشكلةٍ في التصنيع (فالأيادي ذات الأصابع الماهرة كُلفة تصنيعها عالية جدًّا حاليًّا) والجُزء الأخير يرجع إلى مُشكلةٍ في الخوارزمية؛ فإلى الآن، نحن لا نفهم جيِّدًا كيف ندمج قدرتي الحس والتَّحكم للإمساك والتَّلاعُب بالأشياء المُختلفة والمُتنوِّعة الموجودة في أرجاء المنزل العادي. وهناك العشرات من أنواع القبضات المُختلفة للإمساك بالأشياء الصلبة، والآلاف من مهارات التَّلاعُب المختلفة؛ كمهارة هزِّ العُلبة لإخراج حبتي دواءٍ فقط منها، أو نزع المُلصق من على برطمان مُربى، أو فرد الزُّبدة الجامدة على الخُبز الطَّري، أو إخراج شريطٍ واحدٍ من المكرونة الإسباجتي بشوكة ليرى هل نضجت وجاهزة للأكل أم ليس بعد.

يبدو أنَّ مُشكلتي اللَّمس وتصنيع الأيدي ستُحلَّان بتقنية الطِّباعة الثلاثية الأبعاد التي تُستخدم حاليًّا بالفعل في شركة بوسطن ديناميكس لصناعة بعض الأجزاء الأكثر تعقيدًا في آليِّهم «أطلس». أما مهارات التَّلاعُب في الروبوتات فهي تتقدَّم بسرعةٍ كبيرة، وجزء من الفضل راجع إلى التَّعلُّم المُعزَّز.[20] والخطوة الأخيرة، وهي أن ندمج كل هذا معًا في شيءٍ ما يبدأ بالتَّصرُّف ومُحاكاة المهارات البدنية المُبهرة في أفلام الروبوتات التي نراها، ستأتي على الأرجح من مجال التخزين الكئيب بعض الشيء. شركة واحدة وهي أمازون تُوظِّف مائة ألف موظَّف ليحملوا المُنتجات من رُفوف التَّخزين في المخازن الضَّخمة ويشحنوها إلى العملاء. وفي الفترة ما بين عامي ٢٠١٥ و٢٠١٧، كانت أمازون تُطلق تحدِّيًا سنويًّا يُسمَّى «تحدِّي الالتقاط» لتسريع وتيرة تطوير روبوتات قادرة على أداء تلك المُهمة.[21] ما يزال هناك طريق طويل بعض الشيء أمامنا لقطعه، ولكن حين تُحلُّ المشكلات البحثية الأساسية — ربما خلال عقدٍ — يُمكن لنا أن نشهد انتشارًا واسعًا

للروبوتات ذات القدرات العالية. في بداية الأمر سيعملون في المخازن ثمَّ سيُستخدمون في العديد من الأنشطة التجارية التي تكون المهام والأشياء فيها قابلة للتَّنبؤ إلى حدٍّ ما مثل الزراعة والبناء. وأيضًا رُبَّما سنراهم عما قريبٍ في قطاع التجزئة وهم يقومون بمهامَّ مثل تكديس البضاعة على الأرفف في البقالات أو طي الملابس.

وأوَّل من سيستفيد حقًّا من الروبوتات في المنازل هم كبار السِّن والضُّعفاء؛ فالروبوتات يمكن أن توفر لهم قدرًا من الاستقلال لم يكن ليُتاح لولا وجودهم. وحتى لو كانت قدرات الروبوت محدودة وفهمُه لما يحدث حوله فهمًا بدائيًّا، فيمكن أن يكون مفيدًا جدًّا رغم ذلك. أما الروبوت كبير الخدم، على الجانب الآخر، الذي يُدير المنزل بهدوءٍ وثقةٍ ويتوقَّع كل ما قد يُفكِّر فيه سيده، فهو حُلم بعيد حاليًّا؛ إذ يتطلَّب مستوًى قريبًا من الذكاء الاصطناعي العام المُضاهي للذكاء البشري.

(١-٥) الذَّكاء على نطاق عالمي

تطوُّر القدرات الأساسية لفهم الكلام والنُّصوص سيُمكِّن المُساعد الشَّخصي الذَّكي من إتمام المهام التي يستطيع أي مُساعدٍ بشري تنفيذها (لكنَّ المُساعد الآلي سيُنفِّذ تلك المهام لقاء حفنة بنساتٍ شهريًّا عوضًا عن آلاف الدولارات التي يتقاضاها المساعد البشري كل شهر). وكذلك ستُمكِّن القدرات الأساسية لفهم الكلام والنُّصوص الآلات من تنفيذ المهام التي لا طاقة للبشر بها؛ لا بسبب «عُمق» الفهم، ولكن بسبب «نطاقه». فمثلًا الآلة التي تتمتَّع بقدرات القراءة الأساسية ستتمكَّن من قراءة «جميع ما خطَّته يد البشر على مر التاريخ» قبل حلول وقت الغداء، ثمَّ تبدأ البحث عن شيءٍ آخر لتُنجزه.[22] وبقُدرات التَّعرُّف على الكلام، يُمكن للآلة أن «تستمع إلى جميع الحلقات والفقرات التي أُذيعت عبر المذياع أو التلفزيون» قبل العصر. ولتوضيح الأمر بالمُقارنة، فإذا أردنا أن نطَّلع على جميع الكتب والمطبوعات التي صدرت في الفترة الحالية عالميًّا، فإننا سنحتاج إلى توظيف مائتي ألف بشري بدوامٍ كامل (وذلك إذا تغاضينا عن جميع ما كُتب في الماضي)، وتوظيف ستين ألفًا آخرين للاستماع إلى ما يُذاع حاليًّا.[23]

ومثل ذلك النِّظام، إن استطاع فقط أن يستخلص حقائق بسيطة ثمَّ يعمم كل تلك المعلومات على كل اللُّغات الموجودة، سيكون مصدرًا خارقًا للإجابة على الأسئلة وكشف الأنماط؛ رُبَّما يكون أقوى بكثيرٍ من محركات البحث التي تُقدَّر قيمتها الحالية بقُرابة

التريليون دولار. وستكون قيمته البحثيَّة في مجالات مثل التاريخ وعلم الاجتماع لا تُقدَّر بثمن.

من المُمكن أيضًا بلا شكٍّ أن يقدر ذلك النظام على الاستماع إلى جميع مُكالمات الهاتف في العالم أجمع (وهي مهمة ستحتاج إلى نحو ٢٠ مليون شخصٍ). هناك بعض الوكالات السِّرية التي ستجد هذا الأمر مُفيدًا لها. فقد كان بعضُها يقوم بأنواع بسيطةٍ من الاستماع الآلي على نطاقٍ واسعٍ، مثل تحديد كلماتٍ مفتاحيَّة بعينها في المحادثات، لسنواتٍ عديدة، أما الآن فقد تقدَّمت تقنياتها بحيث صارت تدون المُحادثة كلها وتحولها إلى نصٍّ مقروءٍ يُمكن البحث في طياته.[24] وبالتأكيد تلك النُّصوص المدوَّنة مُفيدة، لكنَّها ليست مفيدة مثل الفهم الفوري لجميع المحادثات ودمج محتواها معًا.

إحدى «القدرات الخارقة» الأخرى المُتاحة للآلات هي أنَّها تستطيع أن «ترى العالم كله في آنٍ واحد». إن الأقمار الصِّناعية تُصوِّر العالم كله يوميًّا بمُتوسِّط دقَّة وضوح يصل إلى خمسين سنتيمترًا لكل بكسل. بمثل هذه الدِّقة، فكلُّ بيتٍ على الأرض أو سفينة في البحر أو سيارة على الطريق أو بقرةٍ أو شجرةٍ في مزرعةٍ تكون ظاهرةً وواضحة. ولفحص كل تلك الصور، سنحتاج إلى ما يربو على ثلاثين مليون موظَّفٍ بدوامٍ كامل،[25] ولهذا فإنَّ الغالبية العظمى من بيانات الأقمار الصِّناعيَّة في الوقت الحالي لم يسبق وأن اطَّلع عليها أيُّ إنسان. يُمكن أن تتولَّى خوارزميات الرؤية الحاسوبية مُهمَّة مُعالجة جميع تلك البيانات لإصدار قاعدة بياناتٍ قابلة للبحث فيها للعالم بأكمله تُحدَّث يوميًّا ويُمكن البحث فيها، بل ويُمكن لتلك الخوارزميات أيضًا العمل على تصوُّراتٍ ونماذج تنبُّؤيةٍ للأنشطة الاقتصادية وتغيُّر الغطاء النباتي وهجرة الحيوانات والبشر، وتأثيرات التَّغيُّر المناخي وهلُمَّ جرًّا. ولهذا نرى شركات الأقمار الصِّناعية مثل شركتي بلنت ودجتل جُلوب، مُنهمكةً في العمل على تحقيق تلك الفكرة لتكون واقعًا مُعاشًا.

ولما كانت إمكانية الاستشعار على مستوًى عالمي قائمةً، فكذلك إمكانية اتِّخاذ القرار على مستوًى عالمي أيضًا. على سبيل المثال، إذا استعملنا البيانات التي تُوفِّرها الأقمار الصناعية العالمية يُمكن لنا أن نُنشئ نماذج مُفصَّلة لإدارة البيئة العالمية والتَّنبؤ بالآثار الاقتصادية والبيئية للتدخُّلات البشرية، وتوفير المدخلات التحليلية الضرورية لأهداف الأمم المُتَّحدة للتنمية المُستدامة.[26] وها نحن نرى الآن نظم تحكُّمٍ في ما يُسمَّى «المدينة الذكية»، والتي تهدف إلى تحسين إدارة المُرور والمواصلات العامة وتجميع القمامة وإصلاح الطُّرق

والإصلاحات البيئية والعديد من المهام الأخرى التي تعود بالنَّفع على المواطنين، وقد تتوسَّع هذه النظم لتشمل الدولة كلها لا مدينةً واحدةً وفقط. وحتى وقتٍ قريبٍ، كان هذا المُستوى من التَّنظيم لا يُمكن تحقيقه إلا عبر منظومةٍ روتينيَّةٍ ضخمة وغير فعالةٍ من الموظَّفين، وعاجلًا أم آجلًا، سيُستبدل بهم آلاتٍ ذات قدراتٍ هائلةٍ تقدر على تولِّي الكثير والكثير من نواحي حياتنا المُشتركة نحن البشر. وإلى جانب هذا، بلا شكٍّ تظلُّ إمكانية اختراق الخصوصية وإحكام القبضة على المُجتمعات عالميًّا أمرًا واردًا، وهذا ما سأتناوله في الفصل القادم.

(٢) متى سنشهدُ وُصول الذكاء الاصطناعي الخارق؟

كثيرًا ما يسألني الناس أن أتوقَّع متى سنشهدُ وُصول الذكاء الاصطناعي الخارق، وعادة ما أرفُض الإجابة على هذا السؤال. ولديَّ ثلاثة أسباب تدفعني إلى ذلك. أولًا: هُناك تاريخ طويل من مثل هذه التوقُّعات التي ثبت خطؤها.[27] على سبيل المثال، في عام ١٩٦٠، كتب هربرت سايمن؛ الاقتصادي الحائز على جائزة نُوبل ورائد الذكاء الاصطناعي يقول: «تقنيًّا ... في غضون عشرين سنةً، ستكون الآلات ذات قُدرةٍ على فعل أي عملٍ يستطيع الإنسان عمله.»[28] وفي عام ١٩٦٧، كتب مارفن مينيسكي؛ المُنظِّم المُشارك لورشة عمل دارتموث التي عقدت في عام ١٩٥٦ والتي انبثق منها مجال الذكاء الاصطناعي يقول: «في غضون جيلٍ واحد، أنا على يقينٍ أنَّ عالم الآلات سيُتقن جميع القُدُرات العقلية، اللهم إلا قليلًا منها. وستكون مُشكلة بناء «ذكاءٍ اصطناعي» قد حُلَّت جوهريًّا.»[29]

السبب الثاني لرفضي التَّنبؤ بتاريخٍ نشهدُ فيه الذكاء الاصطناعي الخارق هو أنِّي لا أرى أي عتبة واضحة أمامنا لنتخطَّاها. إن الآلات في الوقت الحالي تفوق القدرات البشرية في بعض المجالات والتي ستتوسَّع وتتعمَّق، ومن المُحتمَل أن تُؤدِّي بنا إلى نظم معرفةٍ عامةٍ خارقة، ونُظُم بحثٍ طبية حيوية خارقة، وروبوتات ماهرة وحاذقةٍ تتمتع بقدراتٍ خارقة، ونظم تخطيطٍ مُؤسَّسيةٍ خارقة، وهلُمَّ جرًّا؛ كُلُّ ذلك سيحدُث قبل أن يكون لدينا نظم ذكاءٍ اصطناعي خارق وعام بالكامل. وتلك النظم التي تحظى ب «شبه ذكاءٍ خارقٍ» ستبدأ، فُرادى ومُجتمعة، في طرح العديد من المشاكل الشَّبيهة التي يطرحُها أي نظامٍ ذي ذكاءٍ عامٍّ.

أما السبب الثَّالث الذي يمنعُني من التَّنبؤ بموعد ظهور الذكاء الاصطناعي الخارق؛ هو أنَّه بطبيعته لا سبيل للتَّنبؤ به. إن الأمر يتطلب العديد من «الطفرات المفاهيميَّة» كما ذكر جون ماكارثي في مقابلةٍ له عام ١٩٩٧[30] والذي أضاف فيها قائلًا: «ما نُريده هو ١,٧ أينشتاين، و٠,٣ من مشروع مانهاتن، ونُريد أشباه أينشتاين أولًا. أظنُّ أنَّ الأمر سيستغرق من ٥ سنوات إلى ٥٠٠ سنة». في القسم التالي سأشرح ماهيَّة بعضٍ من تلك الطفرات؛ وكيف أنه لا يُمكن التَّنبؤ بها؛ فالأمر يُشبه إلى حدٍّ كبيرٍ اختراع سيلارد للتفاعُل النَّووي المُتسلسل بعد عدَّة ساعاتٍ من إعلان رذرفورد أنَّ هذا الأمر يُعدُّ ضربًا من ضُروب الخيال.

ذات مرةٍ في اجتماعٍ للمنتدى الاقتصادي العالمي عام ٢٠١٥، أجبت على هذا السؤال حول متى قد نشهد حُلُول الذكاء الاصطناعي الخارق. كان ذلك الاجتماع مُنعقدًا في إطار قاعدة تشاتم هاوس؛ وهذا يعني أنَّ هُويَّة الحاضرين في ذلك الاجتماع ستكون سريَّة ولن يُعرف صاحب أي مشاركة أو رأي، وإنما يُكتفى بمضمون المشاركة أو الرأي. وحتى مع ذلك، ونبعًا من حرصٍ زائدٍ لديَّ، بدأت إجابتي بقولي: «إجابتي هذه سريَّة وليست للنشر بأي حالٍ من الأحوال ...» وتوقَّعتُ حينها أنَّنا، إن لم نُبلَ بأيِّ كوراث مُعرقلة، فقد نشهدُ وُصول الذكاء الاصطناعي الخارق في حياة أولادي، الذين كانوا صغارًا جدًّا حينها ومن المُحتمل أن يعيشوا عُمرًا أطول بكثيرٍ من الكثير من الحاضرين في ذلك الاجتماع بفضل التَّقدُّم في العُلوم الطِّبية. لم تمض ساعتَان، حتى نُشِر مقال في جريدة «ذا ديلي تليجراف» يذكُر تعليقاتي بجانب صورٍ لروبوتات شريرة مُدمِّرة في ثورةٍ عارمة. وكان عنوان المقال: «الروبوتات «المختلة» قد تقضي على الجنس البشري كاملًا في غضونِ جيلٍ واحد».

إن توقُّعي، الذي يصل إلى قُرابة الثَّمانين عامًا، يتَّسمُ بالتَّحفُّظ أكثر من أيِّ باحثٍ آخر في مجال الذكاء الاصطناعي. فالاستبيانات الحديثة[31] تُشير إلى أنَّ مُعظم الباحثين النَّشطين يتوقَّعون أن نشهد الذكاء الاصطناعي المُضاهي لذكاء الإنسان في مُنتصف هذا القرن تقريبًا. وخبرتُنا مع الفيزياء النَّووية تُشير إلى أنَّ من الحكمة أن نفترض إمكانية حدوث هذا التَّقدُّم مُبكِّرًا عما هو مُتوقَّع ومن ثمَّ فعلينا أن نتجهَّز تبعًا لذلك. وإذا كُنا بحاجةٍ إلى طفرةٍ مفاهيميةٍ واحدةٍ على غرار فكرة سيلارد للتَّفاعل النَّووي المُتسلسل المستحث بالنيوترونات، فإننا قد نشهد وُصول الذكاء الاصطناعي الخارق بصورةٍ ما قريبًا جدًّا وعلى نحوٍ مُفاجئ. والاحتمالات هي أنَّنا لن نكون مُستعدِّين؛ فإذا ما صمَّمنا آلاتٍ ذات ذكاءٍ اصطناعي خارق ولديها قدْرٌ ما من الاستقلاليَّة، فإننا سنجد أنفُسنا في

وقتٍ قصيرٍ غير قادرين على أن نتحكَّم بها. ومع ذلك، فأنا واثق أنَّ لدَينا مُتَّسع من الوقت لأنَّ هناك العديد من الطَّفرات الكبيرة التي نحتاج إليها اليوم وتحول بيننا وبين الذكاء الاصطناعي الخارق، وليست طفرةً واحدةً فقط.

(٣) الطَّفرات المفاهيمية المُتصوَّرة في المُستقبل

تظلُّ مشكلة بناء ذكاءٍ اصطناعي عام يُضاهي الذكاء البشري بعيدةً كُلَّ البُعد عن الحل. إن الحلَّ ليس في دفع المال من أجل مزيدٍ من المُهندسين، ومزيدٍ من البيانات وأجهزة كمبيوتر أكثر ضخامة. بعضُ علماء المُستقبل يُصدرون تخطيطات تستنبِط النمو الأُسي للقُدرات الحوسبيَّة في المُستقبل استنادًا إلى قانون مور، فتراهُم ينشُرون تواريخ متى ستكون الآلات أقوى من أدمغة الحشرات، أو أدمغة الفئران، أو الدِّماغ البشري، أو أدمغة البشر مُجتمِعين، وهكذا.[32] وأقول لكم إن تلك التخطيطات لا فائدة منها؛ فقد قُلتُ سابقًا إن الآلات السريعة تُعطيك الإجابة الخاطئة بسُرعة ليس إلا. وإذا همَّ شخص ما بجمع خُبراء الذكاء الاصطناعي معًا في فريقٍ واحدٍ وأتاح لهُم موارد غير محدودةٍ وأعطى لهم هدفًا واحدًا، وهو تصميمُ نظام ذكاءٍ اصطناعي مُتكامل يُضاهي الذكاء البشري عبر دمج أفضل الأفكار معًا؛ فالنتيجة ستكون الفشل. وسيخرُج النِّظام الذي صمَّمُوه إلى العالم الواقعي ويفشل؛ فلن يفهم ما الذي يحدُث حوله ولن يقدر على التنبُّؤ بعواقب أفعاله، ولن يستطيع فهم ما الذي يُريدُه الناسُ في مواقف الحياة المُتعدِّدة؛ ومن ثَمَّ سيقوم بالكثير من التَّصرُّفات الغبية إلى حدِّ السخافة.

بفهم «كيف» سيفشل هذا النِّظام، يستطيع باحثو الذكاء الاصطناعي أن يتعرَّفوا على المشاكل التي عليهم حلُّها؛ أي الطَّفرات المفاهيمية التي يحتاجون إليها، للوُصُول بمستوى الذكاء الاصطناعي إلى مُضاهاة الذكاء البشري. وسأُبيِّن لكم الآن بعضًا من تلك المشاكل المُتبقِّية والتي إن حُلَّت، رُبما سنجد مشاكل أخرى، لكنَّها لن تكون كثيرةً ومُضنية.

(٣-١) اللُّغة والبديهة

ذكاء من غير معرفة، كمُحرِّك من غير وقود. البشرُ يكتسبون كمًّا هائلًا من المعرفة من أقرانهم من البشر؛ فالمعرفة تنتقل من جيلٍ إلى آخر عن طريق اللُّغة. بعض من تلك المعرفة عبارة عن حقائق: «أوباما انتُخب رئيسًا في عام ٢٠٠٩»، «كثافة النحاس

تبلغ ٨٫٩٢ جرامات لكل سنتيمتر مكعَّب»، «قانون أور-نامو وضع عقوباتٍ للعديد من الجرائم»، وهلمَّ جرًّا. وهكذا فإننا نرى أنَّ قدرًا كبيرًا من المعرفة يكمُن في اللغة نفسها؛ في المفاهيم التي تُتيحُها. فكلمات مثل «رئيس» و«٢٠٠٩» و«كثافة» و«النُّحاس» و«جرام» و«سنتيمتر» و«الجرائم» وبقيَّة كلمات اللغة جميعها تحمل في طياتها قدرًا كبيرًا من المعلومات التي تُمثِّل خُلاصة عمليات الاكتشاف والتَّنظيم التي جعلت تلك الكلمات جزءًا من اللغة في المقام الأول.

لنأخذ على سبيل المثال كلمة «النُّحاس» التي تُشير إلى مجموعةٍ من الذَّرات في الكون، ثُمَّ نقارنها بكلمة «أرجلبرجليوم»؛ وهي كلمة عشوائية وضعتها من مُخيِّلتي لتسمية عددٍ من ذرات الكون اختيرت عشوائيًّا وتُساوي في كمِّيتها عدد ذرات النُّحاس. الواحدُ منَّا يُمكن أن يكتشف العديد من القوانين العامة والمُفيدة والتنبؤية حول النُّحاس؛ كثافتها وقدرتها على التوصيل، وقابليتها للطَّرْق، ودرجة حرارة انصهارها، وأصلها النَّجمي، ومُركَّباتها الكيمائية واستخداماتها العملية وهلُمَّ جرًّا. وفي المُقابل، لا يُمكننا قول أي شيء نهائيًّا حول مادة «أرجلبرجليوم». فالكائن الحي الذي يتحدَّث لُغةً تحتوي على كلماتٍ لا معنى لها مثل «أرجلبرجليوم» لن يكون قادرًا على أداء وظيفته، لأنَّه لن يكتشف أبدًا حالات الانتظام التي تجعله قادرًا على تخيُّل عالمه والتَّنبُّؤ به.

الآلات التي تفهم لُغات البشر فهمًا حقيقيًّا ستكون مُؤهَّلةً لتلقِّي كمياتٍ هائلةٍ من المعرفة البشرية على نحوٍ سريعٍ يجعلها تجتاز عشرات الآلاف من السنين من التَّعلُّم قضاها أكثر من مائة مليار إنسانٍ عاشوا على وجه الأرض. فمِن غير العملي أن نتوقع أن تُعيد الآلات اكتشاف جميع تلك المعرفة من الصفر، بداية من البيانات الحسيَّة الخام.

لكن، في الوقت الحالي، تقنية اللغة الطبيعية لا تقوى على تنفيذ مهمة قراءة وفهم ملايين الكُتب – التي قد يُحيِّر الكثير منها حتى إنسانًا ذا علمٍ وخبرة. إن نظم مثل نظام «واطسون» من تصميم شركة آي بي إم الذي هزم بطلي مُسابقة «المحك» الأمريكية هزيمةً ساحقةً عام ٢٠١١، يُمكنه استخلاص معلوماتٍ بسيطةٍ من حقائق واضحة من النُّصوص، لكنَّه يعجز عن بناء تراكيب معرفيةٍ مُعقَّدةٍ منها؛ ولا يُمكنه أيضًا الإجابة على الأسئلة التي تتطلب تفكيرًا منطقيًّا مُوسَّعًا في معلوماتٍ من مصادر مُتعدِّدة. على سبيل المثال، مهمَّة قراءة جميع الوثائق المُتاحة حتى نهاية عام ١٩٧٣ وتقييم (مع الشَّرح)

النتائج المُتوقَّعة لعملية اتِّهام الرئيس الأمريكي حينها؛ نيكسون، بفضيحة «ووترجيت»، ستكون صعبة التحقيق في ضوء إمكانياتنا الحالية.

هناك العديد من الجهود الجادة التي تهدف حاليًّا إلى تعميق مستوى التحليل اللُّغوي واستخلاص المعلومات. على سبيل المثال، مشروع «أريسطو» بمعهد ألين للذكاء الاصطناعي يهدفُ إلى تصميم نُظُمٍ تستطيع اجتياز اختبارات العلوم المدرسية بعد قراءة المناهج التعليمية والأدلة الدِّراسية.[33] وفيما يلي سؤال من اختبار الصف الرابع:[34]

> طلاب الصَّف الرابع يُنظِّمُون سباقًا بأحذية التَّزلُّج. أي سطحٍ من الأسطح التالية سيكون أفضل لمثل هذا السباق؟
>
> (أ) الأرض الحصباء.
> (ب) الأرض الرملية.
> (ج) الأرض الأسفلتية.
> (د) الأرض المُعشوشبة.

الآلة التي ستُجيب على هذا السؤال ستُواجه مصدري صُعوبة على الأقل. الأول هو المُشكلة التقليدية لفهم اللغة وما يعنيه هذا السؤال؛ تحليل البنية النَّحويَّة، وفهم معاني الكلمات وما إلى ذلك. (جرِّب ذلك بنفسك: استخدم أي موقعٍ من مواقع الترجمة الآلية المُتاحة على الإنترنت لتُترجم هذا السؤال إلى لُغةٍ تجهلُها، ثمَّ استخدم قاموسًا لتلك اللُّغة وحاول ترجمتها عكسيًّا إلى اللغة الأصلية.) أما مصدر الصُّعوبة الثاني، فهو الحاجة إلى بديهةٍ وإدراكٍ عام لكي تفهم الآلة أنَّ هذا السباق هو على الأرجح سباق بين أُناسٍ يرتدون أحذية التَّزلُّج (في أقدامهم)؛ وأنَّ «السَّطح» هو ما سيتزلَّج عليه المتزلجون وليس ما سيجلس عليه المُشجِّعون؛ وأنَّ كلمة «أفضل» تصف في هذا السِّياق سطح أرض السِّباق وهكذا دواليك. تخيَّل كيف قد تكون الإجابة إذا غيَّرنا عبارة «طُلاب الصَّف الرابع» إلى عبارة «مُدرِّبو مركز تدريبٍ عسكري ساديون».

وإذا أردنا أن نُلخِّص صعوبة الأمر فيُمكن لنا أن نقول إن القراءة تحتاج إلى معرفةٍ والمعرفةُ (في مُعظمها) تأتي من القراءة. بعبارةٍ أخرى، نحن نُواجه مُعضلة الدجاجة والبيضة الكلاسيكية. قد نأمل إذن في وجود عملية تمهيدٍ ذاتي تُمكِّن النظام من قراءة بعض النصوص السهلة واكتساب بعض المعرفة منها، ثمَّ استخدام تلك المعرفة لقراءة

نُصوصٍ أصعب ليكتسب معرفةً أكثر وهكذا دواليك. ولسوء الحظِّ، ما يحدُث غالبًا هو العكس؛ فالمعرفة المُكتسبة معظمها خاطئ؛ ومن ثمَّ تُسبِّب أخطاء في القراءة التي تؤدِّي بدورها لمعرفةٍ أكثر خطأً وتستمرُّ الدائرة.

على سبيل المثال، مشروع «نيل» (مشروع تعلُّم اللغة المُستمر) بجامعة كارنيجي ميلون الذي ربما يُعد أكثر المشاريع الواعدة في تعلُّم اللغات في الوقت الحالي باستخدام عملية التمهيد الذاتي. منذ عام ٢٠١٠ إلى عام ٢٠١٨، اكتسب المشروع ما يزيد على ١٢٠ مليون مُعتقد عبر قراءة النُّصوص الإنجليزية على الويب.[35] بعض تلك المُعتقدات صحيح تمامًا، مثل أنَّ مابيل ليفز يلعب الهوكي وفاز بكأس ستانلي. وإلى جانب الحقائق، يكتسب «نيل» طوال الوقت مُفرداتٍ وتصنيفاتٍ وعلاقاتٍ دلالية جديدة. وللأسف، فإن «نيل» لديه ثقة فيما يُقارب ٣ بالمائة فقط من مُعتقداته ويعتمد على الخُبراء البشريين لحذف المُعتقدات الخاطئة أو التي لا معنى لها من ذاكرته على نحوٍ دوري، ومن أمثلة هذه المُعتقدات أنَّ ««نيبال» عبارة عن «دولة» تُعرف أيضًا باسم «الولايات المُتَّحدة»»، وأن ««القيمة» هي «مُنتج زراعي» غالبًا ما يُقسَّم إلى «أساس»»».

أنا أظُنُّ أنه لا تُوجَد طفرة علمية واحدة يُمكن أن تقلب الدُّنيا رأسًا على عقب. فعملية التمهيد الذاتي الأساسية تبدو صحيحة؛ البرنامج الذي يعرف عددًا كافيًا من الحقائق يستطيع أن يتعرَّف على الحقيقة التي تُشير إليها جُملةٌ ما جديدة؛ ومن ثمَّ يتعلَّم شكلًا نصيًّا جديدًا للتعبير عن الحقائق يُتيح له التَّعرُّف على مزيدٍ من الحقائق، وتستمر العملية هكذا. (نشر سيرجي برن؛ الشَّريك المُؤسِّس لشركة جوجل، بحثًا مُهمًّا عن فكرة التمهيد الذاتي عام ١٩٩٨.)[36] فإذا بدأنا العمل بضخِّ قدرٍ كافٍ من المعرفة المُشفَّرة يدويًّا والمعلومات اللُّغوية، فسيُساعد هذا في تحريك عجلة التَّقدُّم بلا أدنى شك. وزيادة تعقُّد تمثيل الحقائق — مما يسمح بأحداثٍ مُعقَّدة وعلاقات عابرة ومُعتقدات ومواقف الآخرين، وما إلى ذلك — وتحسين التَّعامُل مع عدم اليقين فيما يتعلَّق بمعاني الكلمات ومعاني الجُمَل قد يؤديان في النهاية إلى عملية تعلُّمٍ ذاتي التعزيز عوضًا عن ذاتي الانطفاء.

(٣-٢) التعلم التراكمي للنظريات والمفاهيم

قبل قُرابة ١,٤ مليار عامٍ وعلى بُعد ٨,٢ سيكستيليون ميلٍ من الأرض، كان هناك ثُقبان أسودان؛ أحدهما كُتلته أكبر من كُتلة كوكب الأرض بـ ١٢ مليون مرَّة، والآخر بـ ١٠ ملايين

مرَّة. اقترب الثُّقبان بما يكفي ليبدأ كلٌّ منهما الدوران حول الآخر، وشيئًا فشيئًا بدآ يفقدان طاقتيهما، ويقتربان أكثر، ويلتفَّان أسرع، حتى وصلا إلى مُعدَّل التفافٍ يُساوي ٢٥٠ مرةً في الثَّانية في دائرةٍ قُطرُها ٣٥٠ كيلومترًا، ثمَّ ما لبثا أن اصطدما ثمَّ اندمجا معًا.[37] وفي اللحظات الأخيرة التي تُقدَّر ببضعة أجزاءٍ من الثانية، كان مُعدَّل الطاقة المُنبعثة في صورة موجات جاذبية أكبر بخمسين مرةً من مجموع الطاقة التي تُنتجها كل نُجوم الكون مُجتمعة. وفي ١٤ سبتمبر عام ٢٠١٥، وصلت تلك الموجات إلى الأرض، وبدأت تمُطُّ وتضغط الفضاء تبادليًّا بمُعدَّل يُقارب ١ لكل ٢٫٥ سيكستيليون ميل، وهو ما يكفي لتغيير المسافة إلى نجم قنطور الأقرب الذي يبعُد ٤٫٤ سنة ضوئية عن الأرض، بمقدار عرض شعرة.

لحُسن الحظِّ، قبل هذه الحادثة بيومَين، كانت الكاشفات المُتطوِّرة لمرصد «ليجو» الأمريكي — والذي يرصُد موجات الجاذبية بمقياس التداخُل الليزري — قُد شُغِّلت في كُلٍّ من واشنطُن ولويزينا. وباستخدام إمكانية قياس التَّداخل بالليزر، كان المرصد قادرًا على قياس التَّشوُّه الطفيف في الفضاء؛ واستنادًا إلى حساباتٍ مبنيةٍ على نظرية النِّسبية العامة لأينشتاين، كان باحثو مرصد «ليجو» قد تنبَّئوا بالشكل الدقيق لموجات الجاذبية المتوقع أن تنتج عن ذاك الحدث العظيم (ومن ثمَّ كانوا يبحثُون عنه).[38]

كان ذلك مُمكنًا بسبب تراكم المعرفة والمفاهيم وتبادلهما بين آلاف البشر عبر قرونٍ من البحث والمُشاهدة. بدايةً من طاليس الملطي الذي كان يفرُك حجر الكهرمان بالصُّوف ويُراقب شُحنة الكهرباء الساكنة وهي تنتُج، ومُرورًا بالعالم جاليليو الذي كان يُلقي الأحجار من أعلى بُرج بيزا المائل، وانتهاءً بنيوتن الذي رأى تُفاحةً تسقط من شجرة، وغيرهم الآلاف من الباحثين والمُشاهدين للظواهر الكونية، استطاعت البشرية أن تضع تدريجيًّا طبقة فوق أخرى من المفاهيم والنَّظريات والآلات؛ مثل الكُتلة والسُّرعة المتجهة والتَّسارُع والقوة وقوانين نيوتن للحركة والجاذبية، ومعادلات المدارات، والظواهر الكهربية، والذرات والإلكترونات والحقول الكهربية والحقول المغناطيسية والموجات الكهرومغناطيسيَّة، والنسبية العامة والخاصة، وميكانيكا الكم وأشباه الموصِّلات ووحدات الليزر وأجهزة الكمبيوتر وما إلى آخره.

يُمكننا، من حيث المبدأ، أن نفهم عملية الاكتشاف هذه بأنَّها تحويل لكل البيانات الحسية التي خبرَها البشر جميعًا إلى فرضيَّة شديدة التَّعقيد حول البيانات الحسية التي رصدها علماء مرصد «ليجو» في يوم ١٤ سبتمبر عام ٢٠١٥ بينما هم ينظرون إلى

شاشات أجهزة الكمبيوتر الخاصة بهم. هذه هي ما تُسمَّى بطريقة التَّعلُّم المُستنِدة إلى البيانات استنادًا محضًا؛ فالبيانات هي المُدخلات، والفرضيات هي المُخرجات وما بين هذا وذلك صُندوق أسود. وإذا ما أمكننا تنفيذ هذه الطريقة، فسنشهد ذروة نجاح منهج التَّعلُّم المُتعمِّق الذي شعاره «بيانات أكثر، شبكة أكبر»، ولكن مع الأسف لا يُمكن تنفيذها. والفكرة الوحيدة المعقولة التي لدَينا الآن وتُفسِّر كيف لكيانات ذكيةٍ أن تُحقِّق شيئًا فريدًا كأن تكتشف أنَّ ثُقبَين أسودَين قد اندمجا معًا، هي أنَّ «المعرفة الفيزيائية المُسبقة» لعلماء المرصد إلى جانب بيانات الرَّصد من أجهزتهم مكَّنتهم أن يتوقَّعوا حدوث اندماج الثُّقبَين معًا. أضف على ذلك أنَّ تلك المعرفة المُسبقة ذاتها كانت نتيجةً للتَّعلُّم بمعرفةٍ مُسبقة أخرى، وهكذا حتى بداية التاريخ البشري. ولذلك فنحن لدَينا، تقريبًا، صُورة «تراكُميَّة» عن كيف يمكن أن تستخدم الكيانات الذكية المعرفة كمادة لبناء قدراتها التنبُّؤية.

قُلتُ «تقريبًا» لأنَّ العلم، بلا شكٍّ، قد انعطف إلى مساراتٍ خاطئةٍ في مراتٍ قليلة فيما مضى من القرون، فنراه ينحرف مُؤقَّتًا سعيًا وراء مفاهيم خياليةٍ مثل الأثير المُضيء والفلوجيستون. لكنَّنا نعرف حقيقةً أن الصُّورة التَّراكُميَّة هي ما حدث «بالفعل»، من حيثُ إن جميع العلماء عبر العُصور دوَّنُوا اكتشافاتهم ونظرياتهم في الكُتب والأبحاث العلمية، ثمَّ أتى العُلماء المتأخِّرون ووجدوا بين أيديهم سُبُل المعرفة الصريحة هذه، وليس التجارب الحسية الأصليَّة للأجيال السابقة البائدة. ولأنَّ أعضاء فريق مرصد «ليجو» عُلماء حقًّا، فقد فطنوا إلى أنَّ جميع أجزاء المعرفة التي استخدموها، بما في ذلك نظرية النسبية العامة لأينشتاين، ما تزال في فترةٍ اختباريةٍ (وستبقى هكذا للأبد) و«رُبَّما» يثبُت خطؤها بالتَّجربة في وقتٍ ما. ولكن تبيَّن أنَّ البيانات التي أصدرها المرصد قدَّمت دليلًا دامغًا على صحة نظرية النِّسبية العامة، بالإضافة إلى طرح أدلةٍ أخرى على أنَّ الجرافيتون، وهو جُسيْم افتراضي حامل لقوة الجاذبية، إنما هو جُسَيْم عديم الكُتلة.

نحن بعيدون جدًّا عن تصميم نظم لتعلُّم الآلة لدَيها القدرة على مُضاهاة قُدرة التَّعلُّم والاكتشاف التَّراكُمِيَّين التي يتمتَّع بها المجتمع العلمي — أو حتى العوام من البشر خلال حياتهم — فضلًا عن التَّفوق عليها.[39] وإذا نظرنا إلى نظم التَّعلُّم المُتعمِّق،«د» فسنجد أنها غالبًا ما تكون مُعتمِدة على البيانات؛ وفي أحسن الأحوال، يُمكننا أن ندخل بعض أشكال المعرفة المسبقة الضعيفة جدًّا في بنية الشَّبكة. أما نظم البرمجة الاحتمالية،«ج» فإنها تسمح للمعرفة المُسبقة أن تُوجَد في عملية التَّعلُّم، وتُعبِّر عنها في بنية القاعدة المعرفية الاحتمالية

ومُفرداتها. ومع ذلك، ليس لدينا حتى الآن طرائق فعالة لإنتاج مفاهيم وعلاقاتٍ جديدة، واستخدامها في توسيع القاعدة المعرفية هذه.

لا تحسب أنَّ الصعوبة هنا تكمُن في إيجاد فرضيات تتوافقُ على نحوٍ جيد مع البيانات؛ فمثلًا يُمكن لنظم التَّعلُّم المُتعمِّق أن تجد فرضياتٍ تتماشى مع بيانات الصور على نحوٍ جيد، وقد بنى علماء الذكاء الاصطناعي برمجيات تعلُّمٍ رمزيةً قادرة على اختصار العديد من الاكتشافات التاريخية للقوانين العلمية الكمِّية.[40] إن عملية التَّعلُّم بالنسبة لكيانٍ ذكي مُستقلٍّ تتطلَّب أكثر من ذلك بكثير.

الأمر الأول هو: ما الذي يجب أن يُضمَّن في «البيانات» التي تُستقى منها التَّنبؤات؟ فمثلًا، في تجربة مرصد «ليجو»، كان النَّموذج، الذي استُعمل للتَّنبؤ بمقدار تمدُّد الفضاء وانكماشه عندما وصلت موجات الجاذبية، يأخذ في حُسبانه معلوماتٍ مثل كُتلتي الثُّقبَين الأسودين المُتصادمين، وسُرعة دوران أحدهما حول الآخر وما إلى ذلك، لكنَّه لم يلتفِت إلى بياناتٍ مثل، في أيِّ يوم من أيام الأسبوع كان ذلك الاصطدام بين الثُّقبَين، أو جدول مباريات الدوري المُمتاز للعبة البيسبول. على الجانب الآخر، فإن النَّموذج المُصمَّم ليتنبأ بالحركة المُروريَّة على جسر سان فرانسيسكو–أوكلاند لا محالة سيأخُذ أيام الأسبوع في اعتباره، وسيُراعي بلا شكٍّ جدول مباريات الدوري المُمتاز للعبة البيسبول، وفي الوقت ذاته، سيتجاهل بياناتٍ ككُتلتَي الثَّقبَين الأسودَين المُتصادمَين وسُرعة دورانهما. وبالمثل، فإنَّ البرمجيات التي تتعلَّم كيفية التَّعرُّف على «أنواع» العناصر في الصور تستخدِم البكسلات كمُدخلاتٍ لها، بينما البرمجيات التي تتعلَّم مهارة تقدير «قيمة» القطع الأثرية يجب أن تعرف معلوماتٍ مثل، المادة التي صُنعت منها كل قطعة، ومَن صنعها ومتى، وتاريخ استخدامها وملكيَّتها وما إلى ذلك. قد تتساءل: لماذا كل هذه المعلومات؟ ببساطةٍ، لأننا نحن البشر نعرف ولو قليلًا من المعلومات عن موجات الجاذبية وحركة المُرور والصُّور المرئية وقطع الأثاث. ونحن نستخدم تلك المعرفة لتحديد المُدخلات المطلوبة للتَّنبؤ بمُخرجاتٍ مُحدَّدة. وهذا يُسمَّى بـ «هندسة الخصائص»، وإجادة تنفيذ تلك العملية يتطلَّب فهمًا جيدًا لِما يُراد التَّنبؤ به بالتَّحديد.

بالطَّبع لا يُمكن لآلةٍ ذكيةٍ حقيقية أن تعتمد على مُهندسين بشريِّين من مُهندسي الخصائص والذين سيظهرون بين الحين والآخر ليُخبِروا الآلة أنَّ هناك شيئًا جديدًا لتتعلَّمه ويُساعدوها في تعلُّمه. لذا، سيكون على الآلة أن ترى بنفسها ما قد يُمثِّل مساحة

افتراضٍ معقولةٍ لمشكلة تعلُّمٍ ما. على الأرجح ستفعل ذلك عبر حشد قدرٍ ضخمٍ من المعرفة ذات الصِّلة وبصيغٍ مُتعدِّدة، لكن في الوقت الحالي، كُلُّ ما لدَينا هو بعض الأفكار الأولية غير الناضجة حول كيفية فعل ذلك الأمر.[41] وكتاب نيلسون جودمان المُسمَّى «الحقيقة والخيال والتَّنبؤ» —[42] الذي كُتب عام ١٩٥٤ وربما يُعدُّ واحدًا من أهم الكُتُب في مجال تعلُّم الآلة التي لا تحظى بالتَّقدير المُناسب — يقترح نوعًا من المعرفة يُسمَّى «الافتراض الأعم»، وهذا النَّوع يُساعد على تعريف حُدود مساحة الافتراض المنطقية. في مثال التَّنبُّؤ بالحالة المُروريَّة، قد يكون الافتراض الأعمُّ ذو الصلة هو أن أيًّا من هذه المعلومات: أي يومٍ من أيام الأسبوع هو ذاك؟ أي ساعةٍ في اليوم حينها؟ ما هي الفعاليات المحلية؟ وما هي آخر أخبار الحوادث والإجازات وتأخُّر جداول المواصلات والطَّقس؟ ومتى تُشرق الشمس ومتى تغرُب؟ قد تؤثِّر على حالة الحركة المرورية. (لاحظ هُنا أنَّك تستطيع استنتاج ذلك الافتراض الأعم من شبكتك المعرفيَّة عن هذا العالم دون الحاجة إلى خبيرٍ في المُرور.) ويستطيع أي نظام تعلُّمٍ ذكي أن يُراكم معرفةً من هذا النوع ويستخدمها في المساعدة في صياغة وحلِّ مشكلات تعلُّمٍ جديدة.

الأمر الثاني والذي رُبما يكون أكثر أهميةً، هو الإنتاج التَّراكُمي لمفاهيم جديدة مثل الكُتلة والتَّسارع والشُّحنة والإلكترون وقوة الجاذبية. فبدون تلك المفاهيم، سيُضطرُّ العُلماء (والعامَّة من الناس) إلى تفسير العالَم المُحيط بهم والتَّنبؤ على أساس المُدخلات الإدراكية الأولية. ولكن إذا تتبَّعنا سَير التاريخ، فإننا نرى أنَّ نيوتن استطاع أن يُكمل ما بدأه جاليليو وغيره من تطويرٍ لمفهومي الكُتلة والتَّسارُع، ونجد أنَّ إرنست رذرفورد قد استطاع أن يثبت أنَّ الذرة تتكوَّن من نواة ذات كثافةٍ وذات شُحنةٍ موجبةٍ وتدور حولها الإلكترونات، لأنَّ مفهوم الإلكترون كان قد طُوِّر في أواخر القرن التاسع عشر (على يد العديد من الباحثِين الذين ساهموا بخطواتٍ صغيرةٍ، الواحدة تلو الأخرى). وبلا شكٍّ، فإن جميع الاكتشافات العلمية مبنيَّة على طبقةٍ فوق الأخرى من المفاهيم التي تمتدُّ عبر الزمان وتتخلَّلُها جميع الخبرات البشرية المُكتسبة.

في فلسفة العُلوم، تحديدًا في بواكير القرن العشرين، كان من المُعتاد أن نشهد أي اكتشافٍ لمفهومٍ جديدٍ يُنسَبُ إلى هذه الصِّفات الثلاثة التي لا يُمكن تعريفها: الحدْسُ والتبصُّر والإلهام. إن كل تلك الصِّفات كانت تُعتبَر مُقاومةً لأي تفسيرٍ منطقي أو حسابي. أما عُلماء الذكاء الاصطناعي بمن فيهم هربرت سايمن شخصيًّا،[43] فقد عارضوا وجهة

النَّظر هذه. فببساطة، إذا كانت خوارزميَّة ما لتعلم الآلة يُمكنها البحث في مساحة افتراضات تتضمَّن إمكانية إضافة تعريفاتٍ لمُصطلحاتٍ جديدةٍ لا تُوجد ضمن مُدخلاتها، حينها يُمكن لتلك الخوارزميَّة أن تكتشف مفاهيم جديدة.

ومثالُ ذلك، لنفرض جدلًا أنَّ آليًّا يُحاول تعلُّم قواعد لُعبة الطاولة عبر مُراقبة مبارياتٍ بين اللاعبين البشريِّين. إنه يُلاحظ كيف أنَّهُم يُلقُون النَّردَين، ثُمَّ يلاحظُ أنَّ اللاعبين يُحرِّكون أحيانًا ثلاث قطعٍ أو أربعًا، بدلًا من أن يُحرِّكوا واحدةً أو اثنتَين، وأنَّ هذا يحدث عندما يكون وجه النَّردَين معًا ١-١ أو ٢-٢ أو ٣-٣ أو ٤-٤ أو ٥-٥ أو ٦-٦. فإذا استطاع البرنامج أن يُضيف مفهومًا جديدًا للثُّنائيات، ويُعرِّفه تبعًا للتَّساوي بين وجهي النَّردين، حينها سيكون البرنامج قادرًا على التَّعبير عن نفس النَّظرية التَّنبُّؤية تعبيرًا أكثر دقةً واختصارًا. إنها عملية واضحة ومُباشرة، تستخدم طرائق مثل برمجة المنطق الاستقرائي[44] لإنشاء برامج مُهمَّتُها اقتراح مفاهيم وتعريفاتٍ جديدة للوصول إلى نظريات دقيقة ومُحكمة في الوقت ذاته.

أما في وقتنا الحالي، فنحن نعرف كيف نفعل هذا في الحالات البسيطة نسبيًّا، ولكن في حالات النَّظريات الأشد تعقيدًا، فإن العدد المُحتمَل للمفاهيم الجديدة التي يُمكن أن تُطرح يُصبحُ عددًا هائلًا لا طاقة لنا به. وهذا يجعل التَّقدُّم الحالي في طرق التَّعلُّم المُتعمِّق في مجال الرؤية الحاسوبية أمرًا مُثيرًا للفُضُول والاهتمام. فالشَّبكات المُعمَّقة غالبًا ما تنجح في التَّعرُّف على سماتٍ وسيطة مفيدةٍ مثل العينين والساقين، والخُطوط والزَّوايا، رُغم أنَّها تعمل بخوارزميات تعلُّمٍ شديدة البساطة. وإذا ما استطعنا أن نفهم على نحوٍ أفضل كيفية حدوث هذا الأمر، يُمكننا تطبيق نفس هذا المنهج لتعلُّم مفاهيم جديدة باللُّغات الأكثر تعبيرية التي تحتاجها العُلوم. هذا الأمر بالتَّحديد سينقل البشرية نقلةً نوعيةً وسيكون خُطوةً فارقةً نحو الذكاء الاصطناعي العام.

(٣-٣) اكتشاف الأفعال

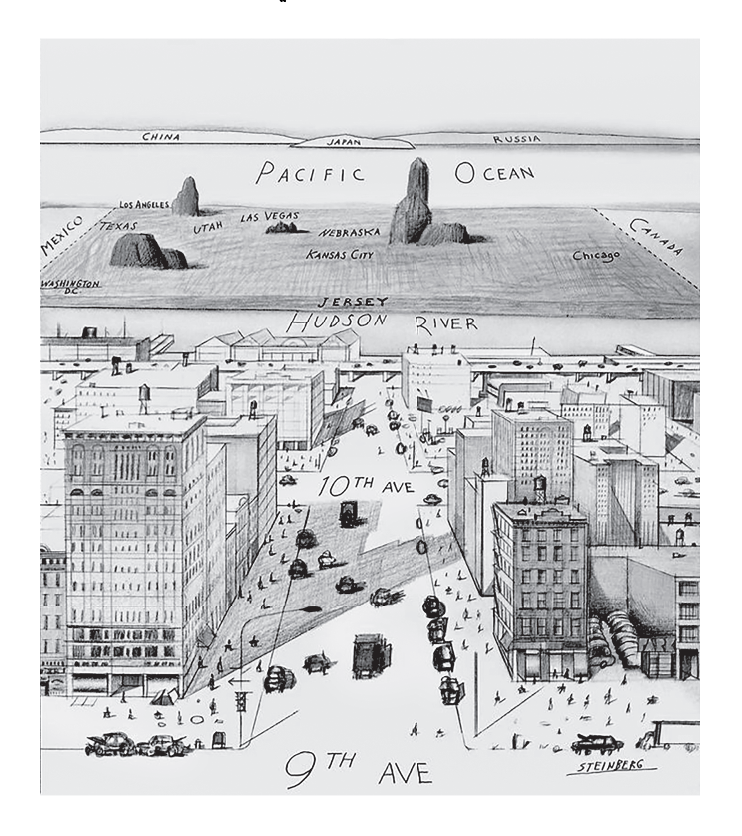
يتطلَّب السُّلوكُ الذَّكي لفتراتٍ طويلةٍ القُدرة على التخطيط للنشاط وإدارته على نحوٍ تسلسُليٍّ؛ وعبر العديد من مُستويات التَّجريد؛ بدايةً مثلًا من تحضير رسالة الدكتوراه (حوالي تريليون فعل)، إلى إرسال أمر تحكُّمٍ حركيٍّ إلى إصبعٍ من أصابع اليد لكتابة حرفٍ واحدٍ في الخطاب التَّقديمي.

أفعالُنا مُرتَّبة في تسلسُلاتٍ هرميَّةٍ مُعقَّدةٍ تنطوي على «عشرات» المستويات من التَّجريد. وتلك المستويات وما تحتويه من أنشطةٍ هي جُزء رئيسيٌّ من حضارتنا البشرية، ويُسلِّمُها جيل إلى آخر عبر وعاء اللغة والمُمارسات العمليَّة. على سبيل المثال، أفعال مثل «صيد خنزيرٍ بريٍّ» أو «التَّقدُّم للحصول على تأشيرة الدخول لبلدٍ ما» و«حجز تذكرة طيرانٍ» قد يتخلَّلُها الملايين من الأفعال البدائية، ومع ذلك فنحن قادرون على التَّفكير فيها كوحدات فردية لأنَّها موجودة بالفعل في «مكتبة» الأفعال التي تُوفِّرها لنا لُغتنا وثقافتنا، ولأنَّنا ندري (بدرجاتٍ مُتفاوتة) كيفية إنجازها.

بمجرد أن تُوجَد تلك الأفعال في المكتبة، فإننا نستطيع أن نشبكها مع أفعال أخرى أكثر تعقيدًا، مثل إقامة مأدبة لأبناء القبيلة بمناسبة الانقلاب الصيفي، أو الشُّروع في بحثٍ أثريٍّ في فصل الصيف بمنطقةٍ نائيةٍ بدولة نيبال. ومُحاولة التَّخطيط لتلك الأنشطة من الصِّفر، بدءًا من خُطوات التحكم الحركيِّ الأكثر بساطة، سيكون ضربًا من العبث، لأنَّ تلك الأنشطة تحوي الملايين أو المليارات من الخُطُوات التي يكون مُعظمها عصِيًّا بشدةٍ على التَّنبؤ. (فأين يا تُرى نجد خنزيرًا بريًّا، وفي أي اتِّجاه سيحاول أن يلُوذ بالفرار؟) أما إذا توفَّرت الأنشطة المعقدة المُناسبة في مكتبتنا، ففي هذه الحالة سنحتاج أن نُخطِّط لبضع خطواتٍ أو نحو ذلك فقط، لأنَّ كُلَّ خطوةٍ من تلك الخطوات ما هي إلا جُزء رئيسيٌّ من أجزاء النَّشاط ككُل. وهذا شيء حتى أدمغتنا الواهنة، نحن البشر، قادرة على التَّعامل معه، لكنه، في الوقت ذاته، يُعطينا «قُوَّةً خارقةً» على التَّخطيط لفتراتٍ زمنيةٍ طويلة.

مرَّ علينا وقت كانت تلك الأنشطة غير موجودة بشكلها الحالي؛ فمثلًا، لتحصل على إذنٍ للقيام برحلة جويَّة في عام ١٩١٠، كان الأمر سيتطلَّبُ خُطواتٍ طويلة وثقيلة على النَّفس وغير مُتوقَّعة من بحثٍ وكتابة خطاباتٍ وتفاوضٍ مع العديد من رُوَّاد الملاحة الجويَّة آنذاك. وتتضمن أنشطة أخرى انضمت مؤخرًا إلى مكتبتنا؛ إرسال رسائل البريد الإلكتروني، والبحث في مُحرِّك البحث «جوجل»، وطلبُ سيارةٍ عبر تطبيق «أوبر». وكما كتب ألفريد نورث وايتهيد في عام ١٩١١ قائلًا: «تتقدَّم الحضارة عندما يزيد عدد الأنشطة المُهمَّة التي يُمكننا فعلها دون الحاجة إلى التَّفكير فيها».[45]

يُظهر الغلاف الشَّهير للرسام سول ستينبيرج لمجلة «ذا نيو يوركر» (انظر الشكل ٣-٢) ببراعةٍ وعلى نحوٍ مكانيٍّ كيف يتحكَّم الكيان الذَّكي في مُستقبله. إن المُستقبل الآني جدًّا شديدُ الوضوح والتَّفصيل؛ في الواقع، كان دماغي قد جهَّز بالفعل سلسلة خطوات



شكل ٣-٢: لوحة «صورة العالم من الجادة التاسعة» للفنان سول ستينبيرج، عام ١٩٧٦. نُشرت هذه اللوحة لأول مرةٍ كغلافٍ لمجلَّة «ذا نيو يوركر».

التحكُّم الحركي المطلوبة لكتابة الكلمات القليلة التالية. وإذا ما نظرت إلى نُقطةٍ أبعدَ في المُستقبل، فسأراها أقلَّ وضوحًا وتفصيلًا؛ فخُطَّتي هي إنهاء هذا القسم من الفصل، ثُمَّ تناول الغداء ثم مُعاودة الكتابة مرَّةً أخرى وبعدها مُشاهدة مُباراة المُنتخبَين الفرنسي والكرواتي في نهائي بطولة كأس العالم لكُرة القدم. وإذا تطلَّع لأبعد من هذا في المُستقبل، فإنَّني سأجد أن خُططي أكبر لكنَّها صارت أكثر غُمُوضًا؛ فأنا أخطط لمُغادرة باريس والعودة إلى بيركلي في أوائل أغسطس، وتدريس مادةٍ لُطلاب الدراسات العُليا وإنهاء هذا الكتاب. وبينما يمرُّ الزَّمانُ، يقترب المُستقبلُ البعيدُ شيئًا فشيئًا من الحاضر وتُصبح

الخُطط أكثر وضوحًا وتفصيلًا، بينما قد تُضاف خُططٌ جديدة وأقل وضوحًا إلى المُستقبل البعيد. أما خُطط المُستقبل الآني، فإنها تكون شديدة الوضوح بشدة حتى إنها لتكون قابلةً للتَّنفيذ مُباشرةً على يد جهاز التحكُّم الحركي.

في الوقت الحالي، لدينا فقط بعض القطع الرَّئيسيَّة للصورة الكلية هذه في مكانها الصَّحيح لبناء نظم ذكاءٍ اصطناعي. وإذا ما توفَّر تسلسُل الأنشطة المُجرَّدة — بما في ذلك معرفة كيفية تحويل كل نشاطٍ مُجرَّدٍ إلى خطة فرعية تتكوَّن من أنشطة ملمُوسة أكثر — حينها سيكون في حوزتنا خوارزميات تستطيع بناء خُططٍ مُعقَّدةٍ لتحقيق أهدافٍ مُحدَّدة. حاليًّا، هناك خوارزميات تستطيع تنفيذ خُططٍ مُجرَّدة ومُتسلسلة هرميًّا بحيث يكون دائمًا لدى الكيان الذكي نشاط بدائي وبدني «جاهز للتنفيذ الفوري»، حتى لو كانت الأنشطة المُستقبلية ما تزال في طور التَّجريد وليست قابلة للتَّنفيذ بعد.

أما القطعة الأساسية المفقودة؛ فهي الوصول لطريقةٍ ما لبناء تسلسُلٍ للأنشطة المُجرَّدة في المقام الأول. على سبيل المثال، هل من المُمكن أن نبدأ من الصِّفر مع روبوت كُلُّ ما يعرفه هو أن بإمكانه إرسال العديد من التيارات الكهربية للعديد من المُحرِّكات ونجعله يكتشف بنفسه فعل الوقوف؟ من المُهمِّ أن أوضِّح أنَّني لا أسأل ما إذا كان بمقدورنا تدريب روبوت على الوقوف أم لا، وهو الأمرُ الذي يُمكننا فعله ببساطة إذا ما طبَّقنا أساليب التَّعلُّم المُعزَّز المربوطة بمُكافأةٍ لدماغ الروبوت عندما يكون جسده بعيدًا عن الأرض.[46] إن تدريب روبوت على الوقوف يتطلَّبُ أن يكون المُدرِّب البشري في الأصل عارفًا بمعنى «الوقوف» ليستطيع تحديد إشارة المكافأة الصحيحة. إن ما نُريده هو أن يكتشف الروبوت بنفسه أن الوقوف هو شيء ما؛ فعل مُجرَّد ومُفيد، وهو شرط أساسي (كونه واقفًا مُنتصبًا على قدمَيه) ليتمكَّن من المشي أو الرَّكض أو المُصافحة بالأيدي أو استراق النَّظر من فوق جدارٍ، وأنَّه من ثمَّ جُزء من العديد من الخُطط المُجرَّدة لتنفيذ جميع أنواع الأهداف. بالمثل، فإننا نُريد من الروبوت أن يكتشف أنشطةً مثل التَّنقُّل من مكانٍ لآخر والتقاط الأشياء وفتح الأبواب وربط العُقد وطهي الطَّعام وإيجاد المفاتيح وبناء المنازل، والعديد من الأنشطة الأخرى التي لا اسم لها في أي لُغةٍ بشرية لأنَّنا نحن البشر لم نكتشفها بعد.

أنا أؤمن أن هذه القُدرة هي أهم خُطوةٍ نحتاجها لبُلوغ الذكاء الاصطناعي المُضاهي لذكاء الإنسان. هذه الخُطوة، بتعبير ألفريد وايتهيد الذي أعيد اقتباس كلامه هنا مرةً أخرى، ستزيد عدد الأنشطة المُهمَّة التي يُمكن لنُظُم الذكاء الاصطناعي فعلها دون الحاجة

إلى التَّفكير فيها. العديد من المجموعات البحثية حول العالَم تبذُل جهدًا جهيدًا لحلِّ تلك المُشكلة. ومن هؤلاء، شركة ديب مايند التي نشرت بحثًا عام ٢٠١٨ يُظهر أداءً يُضاهي مستوى البشر في وضع الإمساك بالعلم في لُعبة «كُويك ٣ أرينا» والتي تدَّعي أنَّ نظم التَّعلُّم لديها «تبني مساحة تمثيلٍ تسلسُلي على نحوٍ مؤقَّت بطريقةٍ جديدةٍ لتعزيز … ترابط سلاسل أنشطة مترابطة على نحوٍ مُؤقَّت.»[47] (أنا لا أدري تحديدًا ما الذي يعنيه هذا الكلام، لكنَّه يبدو بالتأكيد كتقدُّم نحو الهدف المنشُود لابتكار أنشطةٍ معقدة جديدة.) ومع ذلك، أنا لا أظنُّ أنَّ لدَينا حلًّا وافيًا بعدُ لهذه المُشكلة، لكنَّ هذا تقدم قد يحدثُ في أي لحظةٍ، فقط إذا دمجنا بعض الأفكار الحالية معًا بالطريقة الصحيحة.

ستصير الآلات الذكية التي تتمتَّع بهذه القُدرة مؤهَّلةً للنظر إلى مسافةٍ أبعد في المُستقبل والتَّنبؤ به أفضل من البشر. كما سيكون بإمكانها أن تأخذ بعين اعتبارها مزيدًا من المعلومات الهائلة. هاتان القُدرتان معًا ستقودانها لا محالة إلى اتِّخاذ قراراتٍ واقعيةٍ أفضل. وفي أي نوعٍ من أنواع الصراع بين البشر والآلات، سنجد سريعًا، مثل لي سيدول وجاري كاسباروف، أن خُطواتُنا القادمة جميعها قد توقَّعتها الآلات وصدَّتها. وهكذا سنخسر، نحن البشر، الصراع قبل أن يدقَّ طبوله أصلًا.

(٣-٤) إدارة الأنشطة العقلية

إذا كنت تظُنُّ أن إدارة الأنشطة في العالم الواقعي تبدو مُعقَّدةً، فما بالك بإدارة أنشطة «أكثر الأشياء تعقيدًا في هذا الكون»، والذي هو عقلك المسكين؟ إننا نُولد ونحن لا نعرف أي شيءٍ عن كيفية التَّفكير، تمامًا كما لا نعرف أيَّ شيءٍ عن كيفية المشي أو عزف البيانو. إننا نتعلم كيف نُفكِّر. إن بإمكاننا، إلى حدٍّ ما، أن «نختار» أي أفكارٍ نحملُها في دماغنا. (هيا، فكِّر في شطيرة همبُرجر لذيذة ودسمة، أو فكِّر في لوائح نظام الجمارك البلغارية. إنه خيارك!) بطريقة ما، تُعدُّ أنشطتنا العقلية أكثر تعقيدًا من أنشطتنا في العالم الواقعي. وهذا راجع إلى أنَّ أدمغتنا بها أجزاء متحرِّكة أكثر بكثيرٍ من أجسادنا، وتلك الأجزاء تتحرَّك بسرعةٍ فائقة. والأمر ينطبق على أجهزة الكمبيوتر أيضًا؛ فمثلًا لكلِّ تحرُّكٍ من تحرُّكات برنامج «ألفا جو» على رُقعة لُعبة جو، يجرى تنفيذ «ملايين» أو «مليارات» من وحدات الحوسبة، وكُلُّ وحدةٍ من تلك الوحدات تقتضي إضافة فرعٍ لشجرة البحث الاستباقي ثُمَّ تقييم وضع الرُّقعة في نهاية هذا الفرع. وتُنفَّذ كلُّ واحدةٍ من تلك الوحدات لأنَّ البرنامج

يختار أي فرعٍ من فُروع شجرة البحث الاستباقي الذي سيجري استكشافه في الخُطوة القادمة. وعلى نحوٍ تقريبي، فإن «ألفا جو» يختار وحدات الحوسبة التي يتوقع أنَّها ستُحسِّن قراره النهائي في التَّحرُّك على الرُّقعة.

لقد تمكَّنا من وضع نظامٍ مقبول لإدارة نشاط برنامج «ألفا جو» الحوسبيِّ لأنَّ ذلك النَّشاط بسيط ومُتجانس؛ فكُلُّ وحدةٍ من وحدات الحوسبة مثل التي قبلها. وبمُقارنة برنامج «ألفا جو» بالبرامج الأخرى التي تستخدم نفس الوحدة الأساسية للحوسبة، ستجد على الأرجح أنَّه شديد الكفاءة، ولكن إذا ما قُورن بأنواع أخرى من البرمجيات، فربما سنجده عديم الكفاءة بشدَّة. فمثلًا، لي سيدول، الخصم البشري لبرنامج «ألفا جو» في المُباراة التاريخيَّة عام ٢٠١٦، كان على الأرجح لا يُنفِّذ أكثر من بضعة آلافٍ من وحدات الحوسبة في كُلِّ خُطوة، لكنَّه كان لديه هيكل حوسبي أكثر مرونة بكثير، به أنواع مُختلفة من وحدات الحوسبة، بما في ذلك تقسيم الرُّقعة إلى أجزاءٍ فرعية ثُمَّ مُحاولة التَّركيز على كُلِّ جزءٍ على حدةٍ وحلُّه؛ وتمييز الأهداف المُحتملة ووضع خُططٍ معقدة ذات أنشطة مثل «حافظ على هذه المجموعة معًا» أو «صُدَّ الخصم وامنعه من توصيل هاتين المجموعتَين معًا»؛ وأيضًا استبعاد فئات كاملة من التَّحرُّكات لأنَّها تفشل في التعامل مع أحد الأخطار الشَّديدة.

ببساطة، نحن لا نعرف كيفية تنظيم مثل هذه الأنشطة الحوسبيَّة المُعقَّدة والمُختلفة؛ أي كيفية الدمج بين نتائج كُلٍّ منها والبناء عليها، وكيفية تخصيص الموارد الحوسبيَّة للأنواع المُختلفة من التَّفكير والتدبُّر حتى نجد قراراتٍ جيدةً بأسرع ما يُمكن. من الواضح، مع ذلك، أنَّ هيكلًا حوسبيًّا بسيطًا كهذا الذي لدى برنامج «ألفا جو» لا يُمكنه العمل في العالم الحقيقي حيث نحتاج إلى التَّعامل على نحوٍ اعتيادي مع آفاق قراراتٍ تحتوي ليس على العشرات بل المليارات من الخُطوات البدائية، وحيث عدد الأنشطة المُمكنة في أي نقطةٍ هو تقريبًا عدد لا نهائي. من المُهمِّ أن نتذكَّر أن أي كيانٍ ذكي في العالَم الواقعي لن يكون مُقتصرًا على «لعب» لُعبة جو فقط، أو حتى «إيجاد مفاتيحي»؛ فهو يُمكنه فعل «أي شيء» بعد ذلك، لكنَّه لا يُمكنه على الأرجح التَّفكير في جميع الأشياء التي قد يفعلها.

إن أي نظامٍ يُمكنه اكتشاف أفعال معقدة جديدة، كما فصَّلنا سابقًا، بالإضافة إلى إدارة أنشطته الحوسبيَّة للتَّركيز على وحدات الحوسبة التي تُفضي بسرعة إلى تحسُّنٍ كبير لجودة اتِّخاذ القرارات، سيكون صانع قرارٍ لا يُقهَر في العالم الواقعي. وسيضاهي

تفكيره وتدبُّره ما عليه البشرُ من «كفاءةٍ معرفية»، لكنَّه لن يُعاني من الذاكرة الضَّعيفة القصيرة الأمد، أو الإمكانات البطيئة اللتين تُحجِّمان بشدةٍ قدرتنا على استشراف المُستقبل، ومعالجة عددٍ كبير من الأمور الطارئة ووضع عددٍ كبير من الخُطط البديلة.

(٣-٥) أهذا كُلُّ شيء؟

إذا وضعنا معرفتنا عن كلِّ شيءٍ يُمكننا فعله جنبًا إلى جنبٍ مع جميع التَّطوُّرات الجديدة الممكنة المعروضة بين دفَّتي هذا الفصل، فهل سُيجدي هذا نفعًا؟ وكيف سيكون سُلوك النِّظام الناتج؟ ظنِّي أنَّه سيشُقُّ عُباب الزَّمن وسيكتسب كمياتٍ هائلة من المعلومات، وسيُتابع أوضاع العالَم على نطاقٍ واسعٍ عبر المُشاهدة والاستنتاج. وشيئًا فشيئًا، سيُحسِّن من نماذج تصوُّراته عن العالم (بما في ذلك تصوُّراته عن البشر)، وسيستخدم تلك النَّماذج لحلِّ المشاكل المُعقَّدة وسيختزل عمليات الحل ويُعيد استخدامها ليجعل من طريقة تفكيره وتدبُّره طريقةً ذات كفاءة أعلى وليتمكَّن من إيجاد حُلولٍ للمشاكل الأكثر تعقيدًا. وسيكتشف النظام مفاهيم وأنشطة جديدة ستُمكِّنه من تحسين مُعدَّل الاكتشاف لديه، وسيستطيع وضع خُططٍ فعالةٍ لفتراتٍ زمنيةٍ أطول.

خُلاصة القول هي أنَّ من الجليِّ أن لا شيء آخر ذا قيمةٍ كبيرة ينقُصُ هذا الطَّرح، من وجهة نظر النُّظم التي تعمل بكفاءةٍ لتحقيق غاياتها. وبلا شكٍّ، فإن الطريقة الوحيدة لنتأكد من ذلك هي بناء هذا النظام (بعد أن نُحقِّق ما ينقُصُنا من طفراتٍ علميَّةٍ) ثم رؤية ما سيحدُث.

(٤) تخيل كيف هي الآلة ذات الذكاء الخارق

عانى المُجتمع التقني فشلًا ذريعًا في التخيُّل عند مناقشة طبيعة الذكاء الاصطناعي الخارق وتأثيره. إننا غالبًا ما نرى نقاشاتٍ حول تقليل الأخطاء الطبِّية[48] أو حول السيارات الأكثر أمانًا[49] أو حول غيرها من صُور التَّقدُّم ذي الطَّبيعية التَّزايدية. إن الروبوتات يُتخيَّلُون ككياناتٍ فرديَّةٍ تحمل أدمغتها معها، بينما في الواقع قد يكونون غالبًا مُتَّصلين لاسلكيًّا بكيانٍ واحد عام يعتمد على موارد حوسبيَّةٍ ثابتة هائلة. ويبدو الأمر كما لو أنَّ الباحثين خائفون من دراسة العواقب والتَّبعات الواقعيَّة للنجاح في مجال الذكاء الاصطناعي.

إن أي نظام ذكاءٍ اصطناعي عام يُمكنُه، افتراضيًّا، أن يفعل أيَّ شيءٍ يستطيع الإنسان فعله. على سبيل المثال، بعض البشر أجرَوا الكثير من العمليات الرياضيَّة وبذلوا الكثير من الجهد في تصميم الخوارزميات والبرمجة والأبحاث التَّجريبية ليصلوا إلى مُحرِّك البحث الحديث. ولا أحد يُنكرُ أنَّ نتاج كل هذا العمل مُفيد جدًّا وبالطبع قيِّم للغاية. ولكن ما قيمتُه؟ أظهرت دراسة حديثة أن الفرد الأمريكي البالغ العادي من العينة التي أُجريت عليها الدِّراسة يجب أن يُدفع له ١٧٥٠٠ دولارٍ على الأقل نظير أن يتخلَّى عن استخدام مُحرِّكات البحث لمدة عامٍ كامل،[50] مما يعكسُ القيمة العالميَّة لتلك المُحرِّكات التي قد تصل إلى عشرات التريليونات من الدولارات.

والآن تخيَّل معي أنَّ مُحرِّكات البحث غير موجودةٍ بعدُ لأنَّ العمل المطلوب على مدى عقود لاختراعها لم يُنجز، لكن في الوقت ذاته، لدَينا نظام ذكاءٍ اصطناعي خارق. ببساطة، حينها إذا طلبنا من هذا النظام ابتكار مُحرِّكات البحث، فسيكون لدَينا تقنية محرِّكات البحث في غمضة عين، وكل ذلك لأن لدَينا نظام ذكاء اصطناعي خارقًا بين يدينا. سيكون لدينا تقنية بقيمة تريليونات من الدولارات بطلبٍ واحدٍ فقط، ولن نُضطرَّ حتى إلى كتابة سطرٍ واحد إضافي من الشَّفرة البرمجيَّة. قس على ذلك أي اختراع أو سلسلة اختراعات تنقُصُنا؛ فما يُمكن للبشر فعلُه، يُمكن للآلة فعلُه.

هذه النُّقطة الأخيرة تُعطينا حدًّا أدنى مفيدًا (أي تقديرًا مُتشائمًا) لما يُمكن للآلات ذات الذكاء الاصطناعي الخارق فعله. افتراضيًّا، الآلة لدَيها قُدُراتٍ تفوق قُدرة أيِّ إنسانٍ بمُفرده. وهناك أشياءُ كثيرة لا يقدر على فعلها إنسان بمُفرده، لكنَّ جماعةً من البشر عددها «ن» تستطيع تنفيذها، ومثالُ ذلك إرسالُ رائد فضاءٍ إلى القمر، أو صُنع كشاف لموجات الجاذبيَّة، أو اكتشاف تسلسُل الجينوم البشري، أو حُكمُ دولةٍ بها مئات الملايين من الناس. لذا وعلى نحوٍ تقريبي، فإننا سنبني عدد «ن» من نُسخ برنامج الآلة ثُمَّ نُوصل بعضها ببعضٍ بالطريقة ذاتها، مع تزويدها بنفس المعلومات وتدفقات التحكُّم، كما نفعل مع عدد «ن» من البشر. حينها سيكون لدَينا آلة واحدة تستطيع أن تُنفِّذ أي شيءٍ تستطيع مجموعة البشر التي عددها «ن» فعله، بل وبجودةٍ أفضل؛ لأنَّ كُلًّا من المكونات التي عددها «ن» للآلة هو في حدِّ ذاته بمثابة إنسان خارق.

وهذا التَّصميمُ «التعاوني المُتعدِّد الكيانات» لأي نظامٍ ذكيٍّ هو أقلُّ ما يُمكن تصوُّره من القدرات المُمكنة للآلات لأن هناك تصميماتٍ أخرى أكثر كفاءة. في مجموعةٍ من البشر عددها «ن»، إجمالي المعلومات المُتاحة لديهم يظلُّ مُتفرِّقًا بين عدد «ن» من الأدمغة، ويتمُّ

مُشاركته فيما بينها على نحوٍ بطيءٍ ومنقُوص للغاية. ولهذا تُبدِّد المجموعة البشرية التي عددها «ن» مُعظم وقتها في الاجتماعات. في عالم الآلات، لا حاجة لتفريق المعلومات؛ الأمر الذي يُشتِّت الجُهود ويعُوقُ دون رُؤية الصُّورة الكاملة في أغلب الأوقات. ويكفيك قراءةُ سيرة مُختصرة لتاريخ اختراع عقار البنسلين الطويل لتطَّلع على مثالٍ واضحٍ لكيفية تشتُّت الجهود في مجال الاكتشافات العلميَّة.[51]

من الطرق المُفيدة الأخرى لتوسيع خيالك التَّفكير في شكلٍ ما من أشكال المُدخلات الحسِّيَّة، القراءة على سبيل المثال، ثُمَّ توسيع نطاق التفكير. بينما يُمكن للإنسان أن يقرأ ويستوعب كتابًا واحدًا في الأسبوع، يمكن للآلة أن تقرأ وتفهم جميع الكُتُب التي خطَّتها يد البشر، والتي عددها ١٥٠ مليون كتابٍ، في ساعاتٍ قليلة. هذا العمل سيتطلَّبُ كميَّةً لا بأس بها من قُدرة المُعالجة الحاسُوبيَّة، لكنْ يُمكن أن تُقرأ تلك الكتبُ على نحو كبير بالتَّوازي؛ وذلك بإضافة مزيدٍ من الرُّقاقات التي تسمح للآلة أن تُوسِّع من حجم عملية القراءة. ومن نفس المُنطلق، يُمكن للآلة أن ترى كُلَّ شيءٍ في وقتٍ واحد عبر الأقمار الصِّناعيَّة، والروبوتات ومئات الملايين من كاميرات المُراقبة؛ وتُشاهد جميع محطات التِّلفزيون في العالم في وقتٍ واحد؛ وتستمع إلى جميع المحطات الإذاعية والمُكالمات الهاتفيَّة على مُستوى العالم أيضًا. فبسرعة شديدة، ستكون الآلة قد كوَّنت فهمًا مُفصَّلًا ودقيقًا عن العالم وسُكانه، أفضل بكثيرٍ مما قد يطمُح إليه أيُّ إنسان.

يمكن أن يتخيَّل المرء أيضًا أن تتوسَّع الآلات في قدرتها على الفعل. إن الإنسان المُفرد لا يملك أي تحكُّمٍ مُباشرٍ إلا في جسدٍ واحدٍ فقط، بينما الآلة المُفردة يُمكن أن تتحكَّم في الآلاف أو ملايين الآلات الأخرى. والعديد من المصانع المُؤتمتة تستغلُّ هذه الخاصية وتُطبِّقُها بالفعل. أما إذا نظرنا إلى تطبيقات الأمر خارج المصانع، فآلة واحدة يُمكنها أن تتحكَّم بالآلاف من الروبوتات الماهرة لبناء عددٍ كبير من المنازل، على سبيل المثال، يكون كُلُّ منزلٍ فيها مُصمَّمًا ومبنيًّا حسب احتياجات ورغبات سُكانه المُستقبليِّين. أما في المُختبرات، فيُمكن للنظم الآلية الحالية للبحث العلمي أن تُطوِّر قدراتها لتنفيذ ملايين التَّجارب في آنٍ واحد، وربُّما لإنشاء نماذج تنبُّئيةٍ كاملةٍ خاصة بعلم الأحياء البشري يُمكن أن تصل إلى المُستوى الجُزيئي. لاحظ أنَّ قُدرات الآلة الخاصة بالتفكير ستجعلها قادرةً أكثر على اكتشاف نقاط التَّضارُب بين النظريات العلمية، وبين النَّظريات والمُلاحظات. ولا

يُستبعد أنَّنا، في وقتنا الحالي، لدينا ما يكفي من الأدلة التجريبية حول علم الأحياء البشري لوضع علاجٍ لمرض السَّرطان، لكنَّنا لم نُرتِّبها معًا بعْد.

في العالم الإلكتروني، تستطيع الآلات بالفعل الوصول إلى ملياراتٍ من أدوات التَّوجيه؛ وأعني بذلك شاشات كل الهواتف وأجهزة الكمبيوتر في العالم بأسره. وهذا يُفسِّر جزئيًّا قُدرة شركات تكنولوجيا المعلومات على تحقيق ثروةٍ طائلةٍ بعددٍ قليل جدًّا من المُوظَّفين، وهذا الأمر يُشير أيضًا إلى مدى ضعف الجنس البشري وسُرعة تأثُّره بالتَّلاعُب الذي يتعرَّض له عبر الشاشات.

هُناك توسع من نوعٍ آخر يأتي من قُدرة الآلات على استشراف المُستقبل بدقَّةٍ أكبر تفُوق قدرة البشر. لقد رأينا هذا يحدُث بالفعل في لُعبتي الشطرنج وجو، وإذا ما أضيف للآلات قُدُرات مثل وضع وتحليل خُططٍ بعيدة الأمد ذات تسلسُلٍ هرميٍّ؛ واكتشاف أنشطة مجردة جديدةٍ ونماذج وصفيَّة معقدةٍ، فستُنقل هذه الميزة لخدمة مجالات مثل الرياضيات (مما يُؤدِّي لإثبات نظرياتٍ جديدة ومُفيدة)، وعملية اتِّخاذ القرارات في العالم الواقعي. وستكون مهامٌّ مثل إخلاء مدينةٍ كبيرةٍ من سُكانها في حالة إحدى الكوارث البيئيَّة، بسيطةً نسبيًّا؛ فالآلات سيكون بإمكانها إصدار توجيهاتٍ فرديَّة مُخصَّصة لكلِّ شخصٍ ووسيلة نقلٍ لتقليل عدد الضَّحايا.

قد تُضطرُّ الآلات إلى بذل جُهدٍ إضافيٍّ قليل عند محاولة إيجاد اقتراحاتٍ للسياسات العامة للحدِّ من الاحتباس الحراري العالمي. فالتَّخطيط لنُظُمٍ خاصة بكوكب الأرض يتطلَّبُ معرفةً كافيةً بعلم الفيزياء (الغلاف الجوي والمُحيطات)؛ وعلم الكيمياء (دورة الكربون وأنواع التُّربة)؛ وعلم الأحياء (عملية التَّحلُّل والهجرة)؛ والهندسة (الطاقة المُتجدِّدة، واحتباس ثاني أكسيد الكربون)؛ وعلم الاقتصاد (الصِّناعة واستخدامات الطاقة)؛ والطَّبيعة البشرية (الغباء والجشع)؛ والسياسة (غباء أكثر وجشعٌ أكبر). وكما ذكرنا، فالآلات سيكون تحت أيديها كميات ضخمة من الأدلة لتغذية جميع تلك النَّماذج، كما ستكون قادرةً على اقتراح أو تنفيذ تجارب وحملاتٍ استكشافيَّةٍ جديدةٍ للتقليل من حالات عدم اليقين الحتميَّة؛ مثلًا، الوصول إلى الحجم الحقيقي لهيدرات الغاز في خزانات المُحيط الضحلة. كما ستكون الآلات قادرةً على التَّفكير في اقتراحاتٍ للسِّياسة العامة لعددٍ كبيرٍ من المجالات كالقوانين والوكزات بمفهومها السُّلوكي والأسواق والاختراعات وتدخُّلات الهندسة

المناخيَّة، لكنَّها بلا شكٍّ ستحتاج لإيجاد طرق لتُقنعنا بالمُوافقة على تلك الاقتراحات وانتهاجها.

(٥) قيود الذَّكاء الاصطناعي الخارق

لا تسرح بخيالك أكثر من اللازم وأنت تُفكِّر في قُدُرات الذكاء الاصطناعي الخارق. من الأخطاء الشائعة إعطاء الذكاء الاصطناعي الخارق قُدُراتٍ إلهيَّة خارقة من العلم المُطلق والمعرفة غير المحدودة؛ المعرفة الكاملة والمثاليَّة ليست فقط بالحاضر، بل بالمُستقبل أيضًا.[52] وهذا غير مُحتمَل على الإطلاق؛ فهو يتطلَّب قُدرةً غير مادية لتحديد الوضع الحالي الدقيق للعالم، كما يتطلَّبُ قُدرةً لا يُمكن تصوُّر وُجودها لمُحاكاة عمليات العالم الذي تُوجَد فيه الآلات نفسها بسُرعةٍ تسبق وقت حدوثها في الحقيقة (هذا بصرف النَّظر عن مليارات الأدمغة التي ستُعدُّ حينها ثاني أكثر الأشياء تعقيدًا في هذا الكون).

وكلامي هذا لا يعني أنَّ من المُستحيل التَّنبؤ بـ «بعض جوانب» المُستقبل بدرجةٍ مقبولةٍ من اليقين؛ فمثلًا، أنا أعرف أي مادةٍ سأُدرِّس وفي أي قاعةٍ في الجامعة ببيركلي بعد عامٍ تقريبًا من الآن رغم تأكيدات عُلماء نظرية الفوضى بشأن أجنحة الفراشات وتأثيرها وما إلى ذلك. (وأنا لا أعتقد أيضًا أنَّ البشر قد اقتربوا بأي نحو من التَّنبؤ بالمُستقبل في حُدود ما تُتيحُه قوانين الفيزياء!) إن التَّنبؤ بالمُستقبل يعتمدُ على وجود المُجرَّدات الصحيحة؛ فمثلًا، أنا أستطيع أن أتنبأ «أنِّي» سوف أقفُ «على منصَّة قاعة ويلر» في حرم جامعة بيركلي في آخر ثلاثاء من شهر أبريل، لكني لا أستطيع أن أتنبَّأ بموقعي على المنصَّة بدقةٍ قياسًا بالمِلِّيمتر، أو بأيٍّ من ذرات الكربون ستتَّحد مع جسدي في ذلك الوقت.

إن الآلات أيضًا خاضعة لقيود سرعةٍ مُعينة يفرضُها العالم الواقعي على المُعدَّل الذي يمكن من خلاله اكتساب معرفة جديدة عن هذا العالم، وهذه النقطة هي إحدى النِّقاط المُهمة التي أشار إليها كفن كلي في مقاله عن التَّوقُّعات الساذجة عن الذكاء الاصطناعي الخارق.[53] على سبيل المثال، لتحديد ما إذا كان دواءٌ ما يُعالج نوعًا مُعينًا من أنواع مرض السَّرطان في حيوان تجارب، على العالم، سواء أكان بشريًّا أم آليًّا، أن يختار أحد خيارَين: إما أن يحقن الحيوان بالدَّواء ثم ينتظر عدَّة أسابيع، أو يُجري تجربة مُحاكاةٍ دقيقة بشكل كافٍ. ولكن لإجراء مُحاكاةٍ، يتطلَّب الأمر قدرًا كبيرًا من المعرفة التَّجريبية بعلم الأحياء؛ والتي قد لا تتوافر جميعُها في الوقت الحالي؛ لذلك، يجبُ أن يُجرى أولًا مزيدٌ من

التَّجارب الخاصة ببناء النموذج. وبلا أدنى شكٍّ، هذه التَّجارب ستستغرق بعض الوقت ويجب أن تتمَّ في العالم الواقعي.

على الجانب الآخر، يُمكن لعالِمٍ آلي أن يُجري بالتَّوازي عددًا هائلًا من تجارب بناء النموذج، ثُمَّ يدمج نتائج تلك التَّجارب في نموذجٍ مُتَّسقٍ داخليًّا (لكنَّه شديد التَّعقيد)، ثُمَّ يُقارن تنبُّؤات النَّموذج بجميع الأدلة التجريبية المُثبتة في علم الأحياء. زد على ذلك أنَّ مُحاكاة النموذج لا تتطلَّب بالضَّرورة محاكاةً فيزيائية كمية للكائن الحيِّ بالكامل حتى نصِل إلى مستوى التفاعُلات الجزيئية المُفردة. تلك المُحاكاة، كما أوضح كفن كلي، قد تستغرق وقتًا أطول من وقت إجراء التَّجربة في العالم الواقعي. ومثلما أستطيع أن أتنبَّأ، ببعض اليقين، بمكاني المُستقبلي في أيام الثلاثاء من شهر أبريل، يُمكن التنبؤ بدقةٍ بخصائص النُّظُم الأحيائيَّة باستخدام النَّماذج المُجرَّدة. (يرجع هذا، من ضمن أسبابٍ أخرى، إلى أنَّ علم الأحياء يسيرُ على نظم تحكُّمٍ حازمةٍ تعتمد على حلقات التَّقييم المُستمرَّة بحيث لا تؤدي عادة التغيرات الطَّفيفة في الظروف الأولية إلى تغيُّراتٍ كبيرة في النَّتائج.) وهكذا، رغم أن مُساهمة الآلات باكتشافاتٍ «فوريةٍ» في مجال العلوم التَّجريبيَّة تكادُ تكون حُلمًا بعيد المنال، فإنَّنا يُمكن أن نتوقَّع أن العُلُوم سوف تتقدَّم على نحوٍ أسرع بمُساعدتها. وبالفعل هذا واقع نراه بأمِّ أعيُننا في وقتنا الحاضر.

آخر قيود الآلات هو أنَّها ببساطةٍ ليست بشرًا. هذا الأمر يجعلُها في ورطةٍ كبيرةٍ وجوهريَّةٍ عند مُحاولة نمذجة وتوقُّع فئة مُعيَّنة من الأشياء؛ البشر. إن أدمغتنا، نحن البشر، مُتشابهة إلى حدٍّ كبير، ولذلك يُمكن أن نستخدمها لمُحاكاة — أو إن أردنا، معايشة — الحياة العاطفية والفكريَّة للآخرين. وهذا شيء اعتيادي بالنسبة لنا دون أي كُلفةٍ تُذكَر. (إذا أنعمت النَّظر في الأمر، فستجد أن الآلات مُتفوِّقة في هذه النُّقطة فيما بينها؛ فكلٌّ منها يُمكنها فعليًّا تشغيل الشَّفرة البرمجيَّة الخاصة بالآلات الأخرى!) فمثلًا، أنا لستُ بحاجةٍ إلى أن أكون خبيرًا في النُّظُم العصبية الحسِّية لأعرف ما هو شُعور أن تضرب إبهامك بمطرقة؛ فيُمكنُني أن أضرب إبهامي بالمطرقة لأعرف الشُّعور. على الجانب الآخر، على الآلات أن تبدأ تقريبًا[54] من الصِّفر في محاولة فهمها للبشر؛ فكُلُّ ما لدَيها من معلوماتٍ هو عن سُلوكياتنا الخارجيَّة، إلى جانب جميع المراجع والأبحاث في علم النَّفس وعلم الأعصاب، لذلك عليها أن تُطوِّر فهمًا لآلية عملنا، نحن البشر، على ذلك الأساس. من حيث المبدأ، ستقدر الآلات على تحقيق ذلك، لكنَّ من الحكمة أن نفترض أنَّ اكتسابها

لفهمٍ يُضاهي فهم البشر أو يتجاوزه لآلية عمل الإنسان سيستغرق منها وقتًا أطول بكثيرٍ مُقارنة بمعظم القُدُرات الأخرى.

(٦) كيف سينتفع البشرُ بالذَّكاء الاصطناعي؟

ذكاؤنا هو عمادُ حضارتنا. وإذا توصَّلنا إلى ذكاءٍ أعلى، فسيُمكنُ أن نبني حضارةً أعظم ورُبَّما «أفضل» بمراحل كثيرة. تخيَّل أنْ نجد حلًّا لمشاكل كبيرة وعويصةٍ مثل إطالة حياة البشر إلى ما لا نهاية أو اختراع وسائل سفر بسرعةٍ أسرع من الضوء، لكنَّ أحلام الخيال العلمي هذه ليست بعد هي ما يدفعُنا للتَّقدُّم في مجال الذكاء الاصطناعي. (فمع وجود الذكاء الاصطناعي الخارق، سيكون في وسعنا على الأرجح أن نخترع جميع أنواع التقنيات شبه السِّحرية التي كُنا نُفكِّر بها، لكن من الصعب أن نعرف ما قد تكون تلك التقنيات في الوقت الحالي.) لكن لنُفكِّر بدلًا من ذلك في أحد الأهداف الأكثر واقعية بكثير، وهو رفع مُستوى معيشة جميع سُكان الأرض، على نحوٍ مُستدامٍ، إلى مُستوًى يُضاهي مُستويات العيش الكريمة في الدول المُتقدِّمة. باختيار (على نحوٍ اعتباطيٍّ بعض الشيء) أن تعني كلمة «كريمة» المركز المئوي الذي يساوي ٨٨ بالمائة في الولايات المتحدة، فإن ذلك الهدف يُمثِّل زيادةً تُقدَّر بعشرة أضعافٍ تقريبًا في الناتج المحلي الإجمالي عالميًّا، من ٧٦ تريليون دولار إلى ٧٥٠ تريليون دولار سنويًّا.[55]

لحساب القيمة النقدية للعائد من هذا الهدف، يستخدم الاقتصاديون ما يُطلق عليه «صافي القيمة الحالية» لتدفُّق الدخل، والذي يأخذ في الاعتبار خصم الدخل المستقبلي بالنسبة للحاضر. إن للدخل الإضافي الذي يبلغ ٦٧٤ تريليون دولار سنويًّا صافي قيمة حالية تبلغ نحو ١٣٥٠٠ تريليون دولار،[56] بافتراض وجود عامل خصم يبلغ ٥ بالمائة. لذا، ببساطة شديدة، يعدُّ هذا رقمًا تقريبيًّا لما قد تكون عليه قيمة الذكاء الاصطناعي المضاهي للذكاء البشري إن كان بإمكانه تقديم مستوى معيشة كريم للجميع. وفي ظل أرقام كهذه، لا عجب أن الشركات والدول تستثمر مليارات الدولارات سنويًّا في أبحاث وعمليات تطوير الذكاء الاصطناعي.[57] ومع هذا، نجد أن المبالغ المُستثمرة قليلة جدًّا مقارنة بحجم العائد منها.

بالتأكيد كلُّ تلك الأرقام هي مُجرَّدُ توقُّعاتٍ، إلا إذا كان لدى أحدٍ منَّا تصوُّر عن «كيف» يمكن أن يُحقِّق الذكاء الاصطناعي المُضاهي لذكاء الإنسان هذا العمل البُطولي

المتمثل في رفع مُستوى معيشة البشر. إنه يُمكنه فعل هذا بأن يزيد من متوسِّط إنتاج الفرد للسِّلع والخدمات. ولتوضيح الفكرة بعبارةٍ أخرى؛ الإنسان العادي لا يُمكنُه أبدًا أن يتوقَّع استهلاك أكثر مما يُنتجه. ومثالُ سيارات الأجرة الذاتية القيادة الذي ناقشناه فيما سبق من هذا الفصل يُوضِّحُ الأثر المُضاعف للذكاء الاصطناعي؛ ففي ظلِّ الخدمة المُؤتمتة، سيكون من الممكن أن يُدير (لنقُلْ) عشرة رجالٍ أسطولًا كاملًا يحوي ألف مركبة، وهكذا فإنَّ الشَّخص الواحد يُنتِج وسائل مواصلاتٍ أكثر بمائة مرةٍ عن ذي قبل. والأمر نفسه في صناعة السَّيارات واستخراج المواد الأوَّلية الخام التي تُصنع منها السيارات. وبالطَّبْع، بعض عمليات استخراج خام الحديد في شمال أستراليا حيث درجات الحرارة تتجاوز في الغالب ٤٥ درجة مئوية (١١٣ درجة فهرنهايت) قد تمَّت أتمتتُها بالفعل بالكامل في الوقت الحالي.[58]

إن تلك التَّطبيقات الحاليَّة للذكاء الاصطناعي هي نظم مُخصَّصة لأهدافٍ بعينها؛ فالسيارات الذاتية القيادة والمناجم الذاتية التشغيل تطلَّبت استثماراتٍ ضخمة في البحث والتَّصميم الميكانيكي وهندسة البرمجيات وإجراء الاختبارات لتطوير الخوارزميات الضرورية والتأكد من أنَّها تعمل كما ينبغي. تلك هي طريقة إنجاز الأشياء في جميع المجالات الهندسية. وهي أيضًا الطريقة التي كان يتمُّ بها السَّفر أيضًا فيما مضى؛ فإذا كنت تريد أن تسافر من أوروبا إلى أستراليا ثُمَّ العودة مرة أخرى في القرن السابع عشر، فهذا الأمر في حدِّ ذاته يعدُّ مشروعًا ضخمًا سيتكلف مبالغ مالية طائلة ويتطلَّب سنواتٍ من التَّخطيط ويحمل مخاطرةً كبيرة بأن يموت الشخص المسافر. أما الآن فقد اعتدنا على فكرة التَّنقُّل كخدمةٍ مُقدَّمة؛ فإذا أردت أن تكون في مُلبُرن في أوائل الأسبوع القادم، فلن يأخذ الأمر منك سوى عدة نقراتٍ على هاتفك وستدفع مقدارًا ضئيلًا نسبيًّا من المال مُقارنةً بالماضي.

في عصر الذكاء الاصطناعي العام، سيكون «كُلُّ شيءٍ مُقدَّمًا كخدمة». فلن يكون بنا حاجة إلى حشد جيوشٍ من المُتخصِّصين في عُلومٍ مُختلفةٍ، ثُمَّ تنظيمهم في سلاسل هرميةٍ من المتعهِّدين الرئيسيين والفرعيين لتنفيذ مشروعٍ ما. فجميع أشكال الذكاء الاصطناعي العام سيكون لديها وصول لكل معرفة الجنس البشري ومهاراته وأشياء أخرى كثيرة. الفرق الوحيد سيكون في القُدُرات الجسدية؛ فسيكون هناك روبوتات بأرجلٍ وبارعة في استخدام أيديها لعمليات البناء والجراحة، وروبوتات بعجلاتٍ لنقل البضائع على نطاقٍ واسع، وروبوتات على هيئة طوافات رباعية تطوف في السماء لمُهمَّة الفحص الجوي، وهلُمَّ

جرًّا. من حيث المبدأ، وبصرف النظر عن السياسة والاقتصاد، يُمكن لأي شخصٍ أن يكون تحت إمرته مؤسسة كاملة تتكوَّن من الكيانات البرمجيَّة والروبوتات المادية التي تستطيع تصميم وبناء الجُسور، أو تحسين إنتاج محاصيل الأراضي الزراعية، أو طهي العشاء لمائة ضيفٍ، أو تنظيم الانتخابات أو فعل أي شيءٍ آخر يجب فعله. وما يجعل كلَّ هذا مُمكنًا هو «عمومية» الذَّكاء الاصطناعي العام.

أثبت التاريخُ بالطبع أنَّ مُضاعفة الناتج المحلي الإجمالي العالمي للفرد عشر مراتٍ إنما هو أمر مُمكن دون الاستعانة بالذكاء الاصطناعي، لكنَّ الأمر استغرق ١٩٠ عامًا لتحقيقه (من عام ١٨٢٠ إلى ٢٠١٠).[59] تطلَّب الأمر تطوير المصانع والأدوات الآلية والأتمتة والسِّكك الحديد والصُّلب والسيارات والطائرات والكهرباء وإنتاج البترول والغاز الطبيعي والهواتف والمذياع والتِّلفزيون وأجهزة الكمبيوتر والإنترنت والأقمار الصِّناعية والعديد من الاختراعات الثَّورية الأخرى. هذه الزيادة لعشرة أضعافٍ في الناتج المحلي الإجمالي التي ذكرناها في الفقرة السابقة لا يعتمد تحقيقها على مزيدٍ من الاختراعات والتقنيات الثَّورية، بل على قدرة نظم الذكاء الاصطناعي على توظيف ما لدَينا بالفعل من إمكانات في الوقت الحالي ولكن على نحوٍ أكثر كفاءة وعلى نطاقٍ أوسع.

لا شكَّ أننا سنُلاحظ بعض المزايا في حياتنا إلى جانب المنفعة الماديَّة البحتة لرفع مستويات المعيشة. على سبيل المثال، التَّدريس الخُصُوصي معروف أنَّه أكثر كفاءةً بكثير من التَّدريس في الفُصُول، لكن حين يُنفَّذ على يد البشر، فببساطةٍ لا — ولن — يكون مُتاحًا لغالبية الناس. أما مع المُدرِّسين الآليين ذوي الذكاء الاصطناعي، فيُمكن لأي طفلٍ أن يتلقَّى تعليمًا مخصوصًا مهما كان فقيرًا. ستكون تكلفة تعليم الطِّفل الواحد زهيدةً وتكاد لا تُذكَر وسيعيش ذاك الطِّفل حياةً أكثر ثراءً وإنتاجيَّة. وسيغدُو السَّعي وراء الأهداف الفنِّية والفكرية، سواء على مُستوًى فرديٍّ أم جماعي، جزءًا عاديًّا من الحياة بدلًا من أن يكون ضربًا من ضُروب الرفاهية والتَّرف.

أما في المجال الصِّحي، فيُتوقَّع أن تُساعد نظم الذكاء الاصطناعي الباحثين على فهم التَّعقيدات الهائلة لعلم الأحياء البشري والتعامُل معها؛ ومن ثمَّ العمل شيئًا فشيئًا على استئصال جميع الأمراض. وستقُودُنا النَّظرة الأكثر توسُّعًا في علم النَّفس البشري والكيمياء العصبية للبشر إلى إحداث تحسُّنٍ كبير في الصِّحة العقلية.

ربما على نحوٍ غير تقليدي أكثر، يُمكننا أن نتوقَّع أن تُساعد نظم الذكاء الاصطناعي على إيجاد أدوات بناء أكثر كفاءة بكثير للواقع الافتراضي وملء بيئاته بالكثير من الأشياء

الأكثر إثارة بكثير. وهذا قد يُحوِّل الواقع الافتراضي إلى وسطٍ مُحبَّبٍ للتعبير الفنِّي والأدبي، مما يُولِّد تجارب ذات عُمقٍ وثراءٍ لا يُمكنُنا تخيُّلهما في وقتنا الحالي.

أما في الحياة اليومية العادية، فسيتيح المُساعد الذَّكي — إذا صُمِّم على نحوٍ جيد ولم يُلوَّث بالمصالح السياسية والاقتصادية — لجميع الأشخاص إمكانية التَّصرُّف بفعاليَّةٍ بالنِّيابة عنهم في ظلِّ نظامٍ سياسيٍّ واقتصاديٍّ يزداد تعقيدًا، وفي بعض الأحيان عدائيةً، يومًا بعد يومٍ. في الحقيقة، سيكون لديك مُحامٍ، ومُحاسب، ومُستشار سياسي خارق مُستعدون لمُساعدتك في أيِّ وقت. وكما نتوقَّع أن تخفَّ الاختناقات المُرورية عبر دمج ولو عددٍ صغيرٍ من المركبات الذاتية القيادة، يُمكن للمرء منا أن يأمُل في وجود سياساتٍ أكثر رشدًا وصراعاتٍ أقلَّ حدة في ظلِّ بُزوُغ فجرٍ جديدٍ يكون فيه مواطنو العالم أكثر معرفةً وحولهم من ينصحُهم نصائح أكثر حكمةً.

إذا ما حقَّقنا جميع ما ذُكر من تطويراتٍ فقد يُغيِّر ذلك من مجرى التاريخ؛ على الأقل ذلك الجزء من التاريخ الذي كانت تدفعُه الصِّراعات والنِّزاعات بين أبناء المُجتمعات نفسها، وبين بعض المُجتمعات وبعضها، للحُصول على أكبر قطعةٍ من كعكة الحياة. فإذا كانت الكعكة نفسها لا نهائيَّة، فلِمَ إذن الصِّراع مع الآخرين للحُصول على نصيبٍ أكبر؟ سيبدو الأمر كما لو كان الصِّراع على من يحصُل على نُسخٍ رقمية أكثر من جريدةٍ ما؛ فالأمر لا يستحقُّ المعاناة إذا كان أيُّ شخصٍ يستطيع أن يحصُل مجانًا على أي عددٍ يريده من النُّسَخ الرقمية من هذه الجريدة.

تجدُر الإشارة إلى أنَّ هناك حُدودًا لما يُمكن للذكاء الاصطناعي تقديمه. إن كعكتي الأرض والمواد الخام ليستا لا نهائيَّتين، فلا يُمكنُ أن يكون هناك نمو سُكاني لا نهائي، وليس كُل شخصٍ سيكون باستطاعته أن يكون له قصر ذو حديقة خاصَّة. (وهذا سيجعلنا نُفكِّر في التَّعدين في مكانٍ آخر في المجموعة الشَّمسية وإنشاء مُدُنٍ صناعيَّة في الفضاء، لكنني لن أُكمل سردي هذا لأنِّي وعدتُ ألا أتحدَّث حول الخيال العلمي.) وكعكة الفخر ليست لا نهائيَّةً أيضًا؛ ١ بالمائة فقط من الناس يُمكنهم أن يكونوا في طبقة الـ ١ بالمائة التي في القمة. لو كانت السعادة الإنسانية تتطلب الوجود في طبقة الـ ١ بالمائة التي في القمة، فإن الـ ٩٩ بالمائة المتبقِّين من البشر سيكونون حزانى، حتى عندما تكون نسبة الواحد بالمائة المُعدمة الموجودة في القاع تعيشُ حياةً رغدةً ومُرفَّهة.[60] سيكون من المُهمِّ

حينها أن تُقلِّل ثقافاتنا تدريجيًّا من قيمة الفخر والحسد، بكونهما عنصرَين محوريَّين للتَّقدير الذاتي الملمُوس.

وكما قال نيك بوستروم في خاتمة كتابه «الذكاء الخارق»، النَّجاحُ في مجال الذَّكاء الاصطناعي سيُنتج «مسارًا حضاريًّا يقودُنا، نحن البشر، إلى استعمال تلك الهبة الكونية استعمالًا رحيمًا وعطُوفًا.» فإذا ما فشلنا في الاستفادة من منافع الذكاء الاصطناعي، فلا نلُومنَّ إلا أنفُسنا.

الفصل الرابع

إساءة استخدام الذكاء الاصطناعي

يبدو الاستخدام الرحيم والعطوف لتلك الهبة الكونية من جانب البشر أمرًا رائعًا، لكن علينا أن نضع في حسباننا أيضًا معدل الابتكار السريع في مجال الأعمال غير المشروعة. إن الأشخاص ذوي النوايا الخبيثة يسعون لابتكار طرقٍ جديدة لإساءة استخدام الذكاء الاصطناعي بسرعةٍ شديدة لدرجة أنَّ مادة هذا الفصل على الأرجح ستكون قديمة قبل حتى أن يُجرى نشرُه. أتمنَّى أن تنظر إلى قراءة هذا الفصل ليس على أنها دعوة للإحباط ولكن باعتبارها دعوة للعمل قبل أن يفوت الأوان.

(١) المُراقبة والمُطاردة والتَّحكُّم

(١-١) شتازي المؤتمتة

تُعدُّ وزارة أمن الدولة في ألمانيا الشَّرقية، المشهورة أكثر باسم «شتازي»، على نطاق واسع «واحدةً من أكفأ الأجهزة المخابراتية ووكالات الشرطة السِّرية وأكثرها قمعًا على مر التاريخ».[1] لقد كانت لديها ملفات للغالبية العُظمى من سكان ألمانيا الشرقية، وكانت تُراقب المكالمات الهاتفية وتقرأ رسائل البريد، وتزرع كاميرات خفيَّة في الشُّقق والفنادق. وكانت تكتشف بكفاءة الأنشطة المعارضة وتقضي عليها بلا هُوادةٍ أو رحمة. وكان نهجها المُفضَّل في العمل هو التدمير النفسي عوضًا عن السجن أو الإعدام. ولكن هذا المستوى من التحكم كانت كُلفته باهظة؛ فقد أشارت بعض التقديرات إلى أنَّ أكثر من رُبع البالِغين في سن العمل كانوا مُخبرين يعملون لصالحهم، وأنَّ سجلاتهم الورقية وصل عددها تقريبًا إلى حوالي ٢٠ مليار ورقة،[2] وأصبحت مهمة معالجة كمية المعلومات الضخمة التي ترد إليهم واتِّخاذ ردود أفعالٍ مناسبةٍ لها تتخطَّى طاقة وقدرة أي مؤسسةٍ بشرية.

من البديهي إذن أن تُفكِّر وكالات الاستخبارات في إمكانية استخدام الذكاء الاصطناعي في عملهم. لسنواتٍ عدة، كانوا يطبقون نماذج بسيطة من تقنية الذكاء الاصطناعي، بما في ذلك تقنية التَّعرُّف على الصوت، وتمييز الكلمات والعبارات المفتاحيَّة في الأحاديث والنصوص. وبمرور الوقت، تطوَّرت قدرة نظم الذكاء الاصطناعي على «فهم سياق» ما يقوله الناس أو يفعلونه؛ سواء أكان تواصلًا شفهيًّا أم كتابيًّا، أو بالمُراقبة بالكاميرات. في النظم الحاكمة التي تتبني هذه التقنية لأغراض خاصة بالتحكم، يُمكن تصوُّر الأمر كما لو أنَّ لكلِّ مواطنٍ مخبرًا من مُخبري شتازي يُراقبه على مدار الساعة كل يوم.[3]

حتى في المجالات المدنية في الدول التي يتمتَّع مواطنوها بالحرية نسبيًّا، فإننا نخضع للمراقبة الفعالة على نحوٍ متزايد. فالشركات تجمع وتبيع البيانات الخاصة بمشترياتنا واستخدامنا للإنترنت ولشبكات التواصُل الاجتماعي، واستهلاكنا للأجهزة الكهربية وسجلاتنا الخاصة بالاتصال والمحادثات النصية، وتاريخنا الوظيفي وصحتنا. كما يُمكن معرفة مواقعنا من خلال تتبُّع المُكالمات الهاتفية والسيارات المُتَّصلة بالإنترنت. كما أن الكاميرات تتعرَّف على وجوهنا ونحن نسير في الشوارع. كل هذه البيانات وغيرها الكثير، يُمكن أن تُربط خيوطها معًا على يد نظم تكامل المعلومات الذكية لإصدار صورةٍ كاملةٍ إلى حدٍّ ما عما يفعله كل واحدٍ منا، وكيف نعيش حياتنا ومن نُحب ومن نكره، ومن سنُصوِّت له في الانتخابات.[4] وستتفوَّق تلك النظم، حتى إن شتازي الألمانية ستصير مُجرَّد نظامٍ هاوٍ إذا ما قُورنت بها.

(١-٢) التَّحكُّم في سُلوكك

بمجرد أن تُصبح إمكانات المُراقبة جاهزة للاستخدام في تلك النظم؛ فالخطوة القادمة هي تعديل سُلوكك ليتماشى مع أهواء من يُسيِّرُون هذه النُّظُم. ومن الطرق الأولية في هذا الشأن الابتزاز المخصَّص الآلي؛ فالنظام الذي يفهم ما الذي تفعله، سواء بالاستماع إليك أو بقراءة ما تكتبه أو بمُراقبة ما تفعله، يُمكنه بسهولة أن يكتشف الأشياء التي لا يجب عليك فعلها. وإذا وجدك مُتلبِّسًا بشيءٍ ما، فسيتواصل معك للحصول على أكبر قدرٍ من المال منك (أو لإكراهك على القيام بسُلوكٍ ما، إذا كان الهدف هو التَّحكُّم السياسي أو التجسُّس). إن الحصول على هذه الأموال يعمل كإشارة التحفيز المثالية بالنسبة لخوارزميات التَّعلُّم المُتعمِّق، لذلك من المُتوقَّع أن تتطور نظم الذكاء الاصطناعي تطورًا سريعًا في قدرتها على

التَّعرُّف على السُّلوكيات الخاطئة والتَّربُّح منها. في أوائل عام ٢٠١٥، أشرتُ إلى خبير أمنٍ حاسوبي أنَّ نُظُم الابتزاز الآلي المبنية على أساس التَّعلُّم المعزَّز قد تُصبح عما قريبٍ شيئًا واقعيًّا؛ حينها ضحك هذا الخبير وقال لي إن هذه النظم موجودة بالفعل. وأول برنامج ابتزازٍ عُرف وذاع صيته كان يُسمَّى «دليلة»، والذي اكتُشف في يوليو من عام ٢٠١٦.[5]

هناك طريقة أبرع لتغيير سلوك الناس وهي تعديل بيئتهم المعلوماتية بحيث يؤمنون بأشياء مختلفة ويتَّخذون قراراتٍ مختلفة. يستخدم بالطبع المعلنون هذه الطريقة منذ قرون كوسيلة لتغيير سلوك الشراء عند الأفراد. كما أنَّ لحملات الدعاية المُنظَّمة التي هي أداة من أدوات الحرب والهيمنة السياسية، تاريخ أطول بكثير.

إذن، ما الذي اختلف الآن؟ بادئ ذي بدء، لأنَّ أجهزة الذكاء الاصطناعي تستطيع تتبُّع عادات القراءة الإلكترونية لشخصٍ مُعيَّن، وتفضيلاته ومستوى معرفته المُحتمل، فيُمكنها أن ترسِل رسائل مُوجَّهة ومخصصة لزيادة التأثير على ذلك الفرد بينما تُقلِّل من مخاطر إنكار المعلومات الواردة فيها. ثانيًا: نظام الذكاء الاصطناعي سيعرف ما إذا قرأ الشخص الرسالة أم لا، وما المُدة التي قضاها في القراءة وما إذا نقر على أي روابط إضافية مُرفقة في الرسالة أم لا. بعد ذلك، سيستخدم كل هذه الإشارات كتقييمٍ فوري لنجاح أو فشل محاولته للتأثير على هذا الفرد؛ بهذه الطريقة، سيتعلم بسرعة كيف يكون فعالًا أكثر في عمله. وبهذه الطريقة، استطاعت خوارزميات انتقاء المُحتوى على مواقع التواصل الاجتماعي أن يكون لها مثل هذا التأثير الخبيث على آراء المُستخدِمين السياسية.

تغيير آخرُ جديد يتمثل في أن دمج تقنيات الذكاء الاصطناعي والرسوم الحاسوبية وتوليف الكلام، يجعل من الممكن إنتاج ما يُسمَّى ﺑ «التزييف المُتعمِّق»؛ وهو عبارة عن مُحتوًى حقيقي من مشاهد مرئيةٍ ومسموعةٍ لأيِّ شخصٍ وهو يقول أو يفعل أي شيءٍ تقريبًا. هذه التقنية ستتطلَّب أكثر بقليلٍ من مجرَّد وصفٍ شفهي للحدث المُراد تزييفه، مما يجعلها طوع أي شخصٍ في العالم تقريبًا. هل تُريد مقطعًا مُصوَّرًا بالهاتف للسيناتور «س» وهو يتلقَّى رشوةً من تاجر المخدرات «ص» في المؤسسة المشبوهة «ع»؟ بسيطة! هذه النوعية من المحتوى يُمكن أن تُوجِد إيمانًا راسخًا بأشياء لم تحدُث قط.[6] بالإضافة إلى ذلك، تستطيع نظم الذكاء الاصطناعي أن تُولِّد الملايين من الهويات الزائفة، والتي تُسمَّى ﺑ «كتائب الإنترنت»، والتي يُمكنها يوميًّا أن تولِّد مليارات التعليقات والتغريدات والتوصيات، وتُبدِّد بذلك جهود البشر العاديِّين لتبادل المعلومات الحقيقية.

الأسواق الإلكترونية مثل «إي باي» و«تاوباو» و«أمازون»، والتي تعتمد على نُظُم السُّمعة[7] لبناء الثقة بين المشترين والبائعين، هي دائمًا في حربٍ مع كتائب الإنترنت المُصمَّمة لإفساد عملها.

وأخيرًا، وسائل التحكُّم يُمكن أن تكون مباشِرةً إذا استطاعت حكومةٌ ما أن تُفعِّل نظام الثواب والعقاب بناءً على السلوك. إن مثل هذا النظام سيُعامل الناس باعتبارهم خوارزميات تعلُّمٍ مُعزَّز، ويُدربهم على التحقيق الأمثل للهدف الذي وضعته الدولة لهم. والإغراء في هذا الأمر بالنسبة إلى الحكومات، خصوصًا تلك التي لها أسلوب أوتوقراطي، هي أن تُفكِّر كما يلي: سيكون من الأفضل لو تصرَّف الجميع تصرُّفًا جيدًا وتحلَّوا بحِسٍّ وطني وساهموا في تقدُّم الدولة؛ وبما أن التقنية تُساعد على قياس سلوك الأفراد وتصرفاتهم ومساهماتهم، إذن، فسيكون من الأفضل أن نبني نظامًا تقنيًّا للمراقبة والتحكُّم يكون مبنيًّا على مبدأ الثواب والعقاب.

هناك العديد من المشاكل في هذا التفكير. أولًا: هذا التفكير يتجاهل الكُلفة النفسية الناتجة عن العيش تحت نظامٍ قائم على المُراقبة الشديدة والإكراه؛ فالتناغُم الخارجي الذي يُخفي وراءه بؤسًا داخليًّا لا يُمكن أن يُعدَّ وضعًا مثاليًّا أبدًا. إن جميع الفعال الطيبة لن تُصبح كذلك، ولكن ستصير فعالًا لتكثير مجموع النقاط الخاصة بالفرد، وسينظر إليها المُتلقِّي على هذا الأساس. أو الأسوأ من هذا أن مفهوم العمل التطوعي سيختفي تدريجيًّا ليُصبح ذكرى باهتة لشيءٍ اعتاد الناس فعله فيما مضى. فتحْت وطأة هذا النظام، لن يكون لزيارة صديقٍ مريضٍ في المستشفى أي أهمية أخلاقيةٍ أو قيمة عاطفية، وستكون مثلها كمثل وقوفك بالسيارة عند الإشارة الحمراء. ثانيًا: هذا التفكير يقع ضحية لنفس نمط الفشل الذي يقع فيه النموذج القياسي للذكاء الاصطناعي من حيث إنه يفترض أن الغاية المُعلنة هي في الواقع الغاية المُضمَرة الحقيقية. في نهاية المطاف، سيسود قانون جودهارت وسيعمل الأفراد في ظلِّه على التحقيق الأمثل للمعايير الرسمية لقياس السلوك الظاهري، تمامًا كما تعلَّمت الجامعات كيفية التحقيق الأمثل لمعايير الجودة التي تستهدفها نظم تصنيف الجامعات عالميًّا بدلًا من أن تبذُل جهدها في تطوير جودتها الحقيقية (تلك التي لا تقيسُها نُظُم التصنيف).[8] وأخيرًا، فإن فرض معايير مُوحَّدة لقياس جودة السلوك يتغافل بدوره عن نقطة مهمةٍ وهي أن المجتمعات الناجحة هي المُجتمعات التي تتكوَّن من طوائف عديدةٍ من الأفراد يُساهم كل واحدٍ منهم لرخائه بالطريقة الخاصة به.

(١-٣) الحق في الأمن العقلي

إذا نظرنا إلى ما أنجزته الحضارة البشرية، فإننا نجد أنَّ التحسُّن التدريجي في الأمن البدني هو أحد أهم إنجازاتها على الإطلاق. فأغلب البشر يعيشون حياتهم اليومية بلا خوفٍ دائم من الإصابة والموت. كما أن المادة الثالثة من الإعلان العالمي لحقوق الإنسان تنُصُّ على أنَّ «الحياة والحرية والأمن الشخصي هي حقٌّ لجميع الأفراد».

هنا أودُّ أن أُضيف أن الأمن العقلي هو حقٌّ للجميع أيضًا؛ فنحن يحقُّ لنا أن نعيش في بيئةٍ تعُمُّها البيانات الحقيقية إلى حدٍّ كبير. إن البشر يميلون إلى تصديق الأدلة التي يرَونها بأعينهم ويسمعونها بآذانهم؛ فنحن نثِق في عائلاتنا وأصدقائنا ومُعلِّمينا و(بعض) المصادر الإعلامية عندما يُخبروننا أنَّ ما يُؤمنون به هو الحق والحقيقة. ورغم أنَّنا لا نتوقَّع أن ما يُخبرنا به بائعو السيارات المُستعملة أو السياسيُّون هو الحقيقة، فإننا نُواجِهُ صعوبةً في تصديق أنهم قد يكذبون وبوقاحةٍ كما يفعلون أحيانًا. ولهذا، فنحن كائنات شديدة الضَّعف في مواجهة التقنية التي تُروِّج للمعلومات المُضلِّلة.

والحق في التَّمتُّع بالأمن العقلي يبدو أنَّه لا يحفل بأيِّ أهميةٍ في الإعلان العالمي. إن المادتَين الثامنة عشرة والتاسعة عشرة تنُصَّان على حقوق «حرية التفكير» و«حرية الرأي والتعبير». وبلا شك، فإنَّ تفكير المرء وآراءه تُبنى ولو جزئيًّا على البيئة المعلوماتية التي يكون فيها؛ ومن ثمَّ فإنها تخضع لنصِّ المادة التاسعة عشرة التي تنصُّ على «الحق في مشاركة المعلومات والأفكار من خلال أي وسيلةٍ إعلامية ودونما اعتبار للحدود الجغرافية.» وهذا يعني أنَّ أيَّ شخصٍ في أي مكانٍ في العالم، لدَيه الحق في نقل المعلومات الزائفة إليك. وهنا مَكمن الصُّعوبة؛ فالأمم الديموقراطية، وعلى وجه الخصوص الولايات المتحدة الأمريكية، كانت ولا تزال في أغلب الوقت غير راغبة في منع تناقُل الأخبار الزائفة في الأمور العامة بسبب المخاوف المُبرَّرة من التحكُّم الحكومي في حرية التعبير (أو غير قادرة دستوريًّا على ذلك). وبدلًا من اتباع الفكرة التي ترى عدم وجود حرية تفكيرٍ دون وصولٍ للمعلومات الحقيقية، فإنَّ الدول الديموقراطية يبدو أنها وثقت على نحوٍ ساذج في الفكرة التي مفادها أنَّ الحقيقة سوف تنتصر في النهاية، وهذه الثقة العمياء هي ما جعلتنا عُرضةً للخطر من غير حماية. ألمانيا تُمثِّل استثناءً في هذا الشأن، فقد مرَّرت مؤخَّرًا قانونًا يُسمَّى «إقرار القانون في شبكات التواصُل الاجتماعي»، والذي يُلزِم منصَّات تقديم المحتوى بحذف أي محتوًى محظور سواء أكان خطابَ كراهيةٍ أو يتضمَّن أخبارًا

كاذبة، لكن هذا القانون قُوبل بموجةٍ عارمةٍ من النقد بكونه قانونًا غير ديموقراطيٍّ وغير عملي.[9]

إذن، في الوقت الحالي لنا أن نتوقَّع أن يظلَّ أمنُنا العقلي تحت الهجوم، ولا حاميَ له إلا الجهود التجارية والتطوُّعية. تلك الجهود تتضمَّن مواقع تقصِّي الحقائق مثل factcheck.org وsnopes.com، ولكن هناك بالطبع مواقع «تقصِّي حقائق» أخرى تُعلن عن الحقائق على أنها أكاذيب وتُروِّج للأكاذيب على أنها حقائق.

أبرز المؤسَّسات التي تتعامل مع المعلومات مثل جوجل وفيسبوك وُضعت تحت ضغوطٍ شديدةٍ في أوروبا والولايات المتحدة الأمريكية من أجل «فعل شيءٍ حيال هذا الأمر». فها نحن نراهم يُجرِّبون بعض الطرائق للإبلاغ عن المحتوى الكاذب ونبذه باستخدام مُراقِبين آليِّين وبشريِّين على حدٍّ سواء، وتوجيه المُستخدِمين إلى المصادر المُوثَّقة التي تُبطل آثار المعلومات الزائفة. في نهاية الأمر، جميع تلك الجهود المبذولة مَبنية على نظم السُّمعة المُتبادلة؛ فالمصادر تُعتبر مصادر موثوقة لأن بعض المصادر الموثوقة أشادت بها على أنها أهل للثقة. وإذا ما انتشر كمٌّ كبير من المعلومات الزائفة، فإن مثل تلك النُّظم يُمكن أن تفشل فشلًا ذريعًا؛ فالمصادر الموثوقة بالفعل يمكن أن تُصبح غير موثوقة والعكس صحيح، وهذا ما يبدو أننا نراه حاصلًا في وقتنا الحاضر مع المصادر الإعلامية الكبيرة في الولايات المتحدة مثل «سي إن إن» و«فوكس نيوز». وبهذا الصَّدد أشار أفيف أوفديا؛ وهو خبير تقني يعمل في مجال مواجهة المعلومات الزائفة، إلى ما يحدث ووصفه بأنَّه: «نهاية عصر المعلومات؛ فشل كارثي في عالم الأفكار.»[10]

وإحدى طرائق حماية عمل نظم السُّمعة هي إدخال مصادر هي أقرب ما تكون إلى الحقيقة الثابتة. إن حقيقةً واحدةً «تمَّ التأكد من صحتها» يمكن لها أن تُبطل أي عددٍ من المصادر التي أصبحت محل ثقةٍ بطريقةٍ أو بأُخرى إذا ما حاولت نشر معلوماتٍ تُناقض تلك الحقيقة المعروفة. في العديد من البلدان، يعمل الكاتب العدل كمصدرٍ للحقيقة الثابتة ليُحافظ على نزاهة المعلومات القانونية والعقارية؛ فغالبًا ما يكون الكُتَّاب العُدول طرفًا مُحايدًا في أي صفقةٍ، كما أنَّهم يجري اعتمادهم من الحكومات أو الجمعيات المهنية. (في مدينة لندن، تُؤدي شركة ورشيبفول كمباني أوف سكرفينارز» هذا الدور منذ عام ١٣٧٣، مما يدلُّ أنَّ هناك ثباتًا ملحوظًا في دور الإخبار بالحقائق.) وإذا وُضعت المعايير الرسمية والمؤهِّلات المهنية وإجراءات الاعتماد لمتقفِّي الحقائق، فإن هذا سيُساعد على الحفاظ على صحة تدفُّقات المعلومات التي نعتمد عليها. إن منظمات مثل مجموعة

«دبليو ثري سي كردبل ويب» و«كردبلتي كوليشان» تهدف إلى تطوير طرائق تقنية وتعتمد على التعهيد الجماعي لتقييم مُقدمي المعلومات مما سيُتيح للمُستخدِمين تصفية المصادر غير الموثوق بها.

أما الطريقة الثانية لحماية نُظُم السُّمعة فهي بفرض تكلفةٍ على تقديم ونشر المعلومات الزائفة. وهكذا، فإن بعض مواقع تقييم الفنادق تقبل فقط المُراجعات بخصوص فندقٍ ما من الأشخاص الذين حجزوا ودفعوا للمبيت في غرفةٍ من إحدى غُرفه، بينما بعض المواقع الأخرى تقبل المُراجعات مِن كُلِّ من هبَّ ودبَّ. ولا يخفى على أحدٍ أنَّ التقييمات على المواقع الأولى ستكون أقلَّ تحيُّزًا بنحوٍ ملحوظ بسبب التَّكلفة المفروضة على المُراجعات المُزيَّفة (وهي دفع ثمن المبيت في إحدى غرف الفندق دون الذهاب إليه أصلًا).[11] تظلُّ العقوبات «النظامية» محل خلافٍ وإثارة للجدل؛ فلا أحد يُريد أن يرى وزارةً للحقيقة، وفي الوقت ذاته، فإن القانون الألماني السابق الذكر يُعاقب منصة تقديم المحتوى فقط، وليس الشخص الذي شارك الأخبار الكاذبة. على الجانب الآخر، ومع ازدياد عدد الدول وعدد الولايات داخل الولايات المتحدة الأمريكية التي تُجرِّم تسجيل المكالمات الهاتفية دون تصريح، فإنه من المُفترض، على الأقل، أن يكون من المُمكن فرض عقوباتٍ على إنشاء تسجيلاتٍ صوتية ومرئية زائفة للأشخاص الحقيقيِّين.

وأخيرًا، هناك حقيقتان أُخرَيان تصُبَّان في صالحنا. الأولى هي أن لا أحد تقريبًا يُريد عمدًا أن يُخدع وأن يتم التلاعُب به. (أنا لا أقصد بذلك أن الآباء دائمًا ما يتحرَّون الحقيقة أيَّما تحرٍّ ويبحثون عن مدى مصداقية أولئك الذين يمدحون ذكاء أطفالهم ولُطفهم، ولكن أقصد أنَّهم أقل عُرضةً للسَّعي وراء الحصول على استحسان أي شخصٍ معروف عنه أنَّه كذوب.) وهذا يعني أنَّ الأشخاص من جميع الاتجاهات السياسية لدَيهم ما يبعثُهم على تبنِّي الأدوات التي تُساعدهم على التفريق بين الحقائق والأكاذيب. أما الحقيقة الثانية، فهي أن لا أحد يُريد أن يُوصَم بالكذب، وعلى وجه الخصوص المنصات الإخبارية. هذا يعني أنَّ مُقدِّمي المعلومات، خصوصًا أولئك الذين يخافون على سُمعتهم، لدَيهم ما يبعثُهم على الانضمام إلى الجمعيات المهنية والامتثال للقواعد السلوكية التي تدعم قول الحقيقة. وبناءً على ذلك، فإن منصات التواصُل الاجتماعي يُمكنها أن تُقدِّم لمُستخدميها خيار مُشاهدة المُحتوى فقط من المصادر ذات السُّمعة الحسنة التي تمتثل إلى مثل تلك القواعد السلوكية وتُخضع نفسها إلى طرفٍ ثالثٍ لمُراجعة واقتفاء الحقائق.

(٢) الأسلحة الفتاكة الذاتيَّة التَّشغيل

تُعرِّف الولايات المُتَّحدة الأمريكية نظم الأسلحة الفتاكة الذاتيَّة التَّشغيل على أنَّها نظم الأسلحة التي «تُحدِّد موقع الأهداف البشرية وتصوِّب وتقضي عليها دون تدخُّلٍ بشري». وقد وُصفت نظم الأسلحة الذاتيَّة التَّشغيل تلك، لسببٍ وجيهٍ، بأنَّها «الاكتشاف الثوري الثالث في مجال الأسلحة»، بعد اختراع البارود والأسلحة النووية.

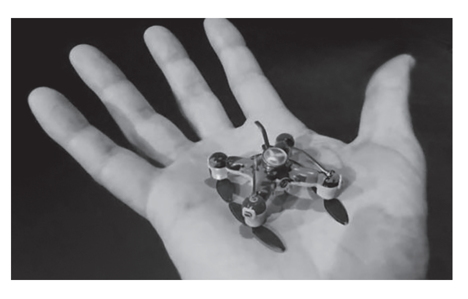
ربما تكون قد قرأت مقالاتٍ في وسائل الإعلام حول نُظم الأسلحة الذاتيَّة التَّشغيل، والتي غالبًا ما ستُطلق عليها «الروبوتات القاتلة» ثمَّ تزيِّن نفسها بصور من سلسلة أفلام «المدمِّر» «ذا تيرمناتر». وهذا الأمر مُضلِّل على الأقل في نُقطتَين؛ الأولى: أنَّه يُصوِّر الأسلحة الذاتية التَّشغيل بأنها خطر مُحدِق لأنَّها قد تسعى للسيطرة على العالم وتدمير الجنس البشري؛ والثانية: أنه يُوحي بأنَّ تلك الأسلحة ستكون على هيئةٍ بشريةٍ ولها وعي، وشريرة.

والتأثير المُجمَل لهذا التَّصوير من وسائل الإعلام لهذه المسألة كان يُحاول تصديرها على أنَّها محض خيالٍ علمي. حتى الحكومة الألمانية انتهجت نفس الطريقة؛ فقد أصدرت مؤخرًا بيانًا[12] تؤكِّد فيه على أنَّ «امتلاك القدرة على التَّعلُّم وتطوير الوعي بالذات يُشكِّل صفةً لا غنى عنها في تعريف المهام الفردية أو نظم الأسلحة بأنها ذاتيَّة التَّشغيل أو مستقلة». (وهذا الكلام يُفهم كما لو أنَّك تُؤكِّد أنَّ الصاروخ لا يُسمَّى ولا يصير صاروخًا إلا إذا تجاوزت سُرعته سرعة الضوء). في الحقيقة، الأسلحة الذاتيَّة التَّشغيل سيكون لها مقدار الاستقلال نفسه الذي يتمتَّع به برنامج لعب الشطرنج، والذي يُعطى مهمة الفوز بالمباراة، لكنَّه يُقرِّر بنفسه تحرُّكاته على رقعة اللعب وأي قطع للخصم سيتخلَّص منها.

الأسلحة الفتاكة الذاتيَّة التَّشغيل ليست خيالًا علميًّا؛ فهي موجودة بالفعل. ورُبَّما أوضح مثالٍ على ذلك هو سلاح الاغتيالات الإسرائيلي «هاروب» (انظر الشكل ٤-١، الصورة التي على اليمين)، وهي طائرة طول جناحيها يُساوي ١٠ أقدام، وبها رأس مُتفجِّرًا يزن ٥٠ رطلًا. وهي تبحث قُرابة ست ساعاتٍ فوق منطقةٍ جُغرافيةٍ مُحدَّدةٍ عن أيِّ أهدافٍ تُوافق المعيار المُحدَّد ثُمَّ تقضي عليها. ذلك المعيار يمكن أن يكون «أي شيء يبُثُّ إشارات رادار ويُشبه الرادار المضادَّ للطائرات» أو «أي شيءٍ يُشبه الدَّبابة».

بدمج الاكتشافات الحديثة في تصميم الطَّوافات الرباعية المُصغَّرة، والكاميرات المُصغَّرة، ورُقاقات الرؤية الحاسوبيَّة، وخوارزميات الملاحة والخرائط، ووسائل اكتشاف





شكل ٤-١: (على اليمين) طائرة «هاروب»؛ سلاح الاغتيالات التي من إنتاج شركة صناعات الفضاء الإسرائيلية؛ (على اليسار) صورة ثابتة من مقطع الفيديو الخاص بالدرون الدقيق «سلوتاربوت» تُوضِّح تصميمًا مُحتملًا لسلاحٍ ذاتي التَّشغيل يحتوي على قذيفةٍ صغيرةٍ مُتفجِّرة.

البشر وتتبُّعهم، فمن المُحتمل أن نرى عما قريبٍ سلاحًا مضادًّا للأفراد مثل الدرون الدقيق «سلوتاربوت»[13] الموضح في الشكل ٤-١ (في الصورة التي على اليسار). مثل هذا السلاح قد يُكلَّف بمهاجمة أي شخصٍ يُوافق معايير بصريةً مُعينة (مثل السِّن والنَّوع والملابس ولون البشرة وهلُمَّ جرًّا)، أو حتى أشخاصًا بعينهم استنادًا إلى تقنية التَّعرُّف على الوجوه. وقد أُخبرتُ أنَّ وزارة الدفاع السويسرية قد بنت بالفعل واختبرت نموذجًا حقيقيًّا من هذا السلاح، وقد وجدت أنَّ تلك التقنية، كما هو متوقَّع، إنما هي تقنية فعالة وعملية وفتاكة في الوقت ذاته.

منذ عام ٢٠١٤ والمحادثات الدبلوماسية جارية في جنيف، وقد تقود إلى معاهدةٍ لحظر الأسلحة الفتاكة الذاتية التَّشغيل. في الوقت نفسه، فإنَّ بعضًا من أبرز المشاركين في تلك المحادثات (الولايات المتحدة الأمريكية، والصين، وروسيا، وإسرائيل والمملكة المتحدة إلى حدٍّ ما) مُنهمكون في منافسةٍ خطيرةٍ لتطوير الأسلحة الذاتية التَّشغيل. على سبيل المثال، في الولايات المتحدة الأمريكية، يهدف برنامج «العمليات المُشتركة في المناطق المتنازع عليها» إلى المُضي قدُمًا نحو الاستقلالية وذاتية التَّشغيل عبر تمكين الدرونات من العمل تحت أقصى ظُروفٍ من انقطاع الاتصالات. ويقول مدير المشروع إن تلك الدرونات «ستصطاد في جماعات كالذئاب».[14] في عام ٢٠١٦، قدَّمت القوات الجوية الأمريكية عرضًا لكيفية نشر ١٠٣ من درونات «بريدكس» الدقيقة من ثلاث طائراتٍ مقاتلاتٍ من طراز «إف-إيه ١٨». وطبقًا للإعلان، «فإن درونات «بريدكس» ليست كيانات فردية مُبرمجة

للتنسيق فيما بينها، بل تعمل كوحدةٍ واحدةٍ وتشارك دماغًا موزعة واحدة لاتخاذ القرارات والتَّكيُّف مع بعضها كأنَّها سرب من الطيور في الطبيعة».[15]

ربَّما تظن أن من الواضح جدًّا أنَّ بناء آلاتٍ يمكنها أن تُقرِّر أن تقتل البشر هو فكرة سيئة جدًّا. لكن عبارة «من الواضح جدًّا» ليست دائمًا مُقنعةً للحكومات، بما في ذلك حكومات بعض الدول المذكورة في الفقرة السابقة، والتي عقدت العزم على تحقيق ما تظنه تفوقًا استراتيجيًّا. سبب آخر أكثر إقناعًا يدعونا لنبذ فكرة الأسلحة الذاتية التَّشغيل هو أنَّها «أسلحة قابلة للتَّوسُّع قادرة على إحداث دمارٍ شامل».

ومصطلح «قابلة للتَّوسُّع» هو أحد مصطلحات مجال علم الكمبيوتر، وتُوصف عملية ما بأنَّها قابلة للتَّوسع إذا كان بإمكانك تنفيذ مليون نسخةٍ منها إذا اشتريت مُكونات كمبيوتر مادية أكثر بمليون مرة. ومثال ذلك هو ما نراه من شركة جوجل التي تُعالج قُرابة الخمس مليارات عملية بحثٍ في اليوم الواحد، ليس بتوظيف ملايين الموظفين، ولكن باستخدام ملايين أجهزة الكمبيوتر. وبشأن الأسلحة الذاتية التَّشغيل، فباستطاعتك أن تُنفِّذ عمليات قتلٍ أكثر بمليون مرة إذا اشتريت أسلحة أكثر بمليون مرة، وهذا راجع تحديدًا إلى أنَّها «أسلحة ذاتية التَّشغيل». فبخلاف الدرونات المُسيَّرة عن بُعد أو رشاشات «آي كيه ٤٧»، فإن تلك الأسلحة الذاتية التَّشغيل لا تحتاج إلى أفرادٍ بشريين لمُراقبة عملهم.

باعتبارها أسلحة دمارٍ شاملٍ، فإن تلك الأسلحة الذاتية التَّشغيل القابلة للتوسُّع تُعطي المُهاجم بعض المُميزات إذا ما قُورنت بالأسلحة النووية والقصف البساطي؛ فهي تترك المباني والأماكن من غير أذًى، ويُمكن أن تُرسَل لتنتقي فقط أولئك الذين قد يُهدِّدون قوات أجنبية مُحتلَّة وتقضي عليهم. وهذا السلاح قد يُستخدَم بالتأكيد لمحو طائفة عرقيةٍ بأكملها من على وجه الأرض أو جميع أتباع دينٍ بعينه (إذا كان لأتباعه صفة ظاهرية مُميِّزة). وفوق كل ذلك، في حين أن استخدام الأسلحة النووية يُعد عتبةً كارثيةٍ، قد نجحْنا (لا لشيءٍ سوى بالحظ المحض) في تجنُّبها منذ عام ١٩٤٥، فإن الأسلحة الذاتية التَّشغيل القابلة للتوسُّع ليس لها مثل تلك العتبة. فالهجمات يُمكن أن تشتد ضراوتها بسلاسةٍ لتصل من ١٠٠ ضحية إلى ١٠٠٠ ضحيةٍ إلى ١٠ آلاف ضحيةٍ إلى ١٠٠ ألف ضحية. وبالإضافة إلى الهجمات الفعلية، فإن مُجرَّد «التهديد» باستخدام هذه الأسلحة يجعلها أداةً فعالة لنشر الرُّعب والقمع. إن تلك الأسلحة ستُقلِّل بشدةٍ من أمن الإنسان على جميع المستويات؛ الشخصي والمحلي والوطني والدولي.

هذا لا يعني أنَّ الأسلحة الذاتية التَّشغيل ستُساهم في نهاية العالَم كما صُوِّر الأمر في سلسلة أفلام «ذا تيرمناتر». إن تلك الأسلحة لا يجب أن تكون ذكيةً على وجه خاص — قد تحتاج السيارات ذاتية القيادة إلى ذكاءٍ أكبر منها — ولن تكون مُهمَّتُها من نوعية المهام «التي تسعى للسيطرة على العالم». إن الخطر الوجودي للذكاء الاصطناعي لن يأتي في المقام الأول من بعض الروبوتات القاتلة ذات الذكاء المحدود. على الجانب الآخر، الآلات ذات الذكاء الخارق إذا تصادمت مع الجنس البشري، فقد تُسلِّح بالطبع نفسها بهذه الطريقة، بتحويل هؤلاء القتلة الآليين الأغبياء نسبيًّا إلى امتداداتٍ مادية لنظام تحكُّمٍ عالمي.

(٣) القضاء على مفهوم العمل الذي عهدناه

الآلاف من المقالات وأعمدة الرأي في الجرائد وغيرها من وسائل الإعلام، والكثير من الكُتب كُتبت حول موضوع استيلاء الروبوتات على وظائف البشر. مراكز الأبحاث تظهر حول العالم لفهم ما الذي سيحدث على الأرجح.[16] ويُلخِّص عنوان بحث مارتن فورد «بُزوع فجر الروبوتات: التقنية وخطر المُستقبل الخالي من الوظائف»،[17] وعنوان بحث كالوم تشيس «التَّفرُّد الاقتصادي: الذكاء الاصطناعي وموت الرأسمالية»[18] القلق حيال هذا الأمر تلخيصًا مُمتازًا. ورغم أنِّي لستُ مؤهلًا بأي حالٍ من الأحوال (كما سيتَّضح لاحقًا) للنقاش في هذه النقطة التي هي في صُلبها أمرًا لعُلماء الاقتصاد،[19] فإني أظن أنَّ هذه المُشكلة شديدة الأهمية بحيث نترك أمرها للاقتصاديين وحدَهم.

مُشكلة «البطالة التقنية» ظهرت لأول مرةٍ في مقالٍ مشهورٍ كتبه جون مايناردكينز تحت عنوان «الخيارات الاقتصادية لأحفادنا». لقد كتب هذا المقال في عام ١٩٣٠ عندما أصاب بريطانيا الكساد الكبير وتسبَّب في موجةٍ عارمةٍ من البطالة، لكنَّ هذا الموضوع له تاريخ أقدم بكثير. لقد قدم أرسطو النقطة الرئيسية بوضوحٍ شديد في الباب الأول من كتابه «السياسة» وقال:

> إذا افترضنا أنَّ كلَّ آلةٍ تقدر على إنجاز عملها، وتُطيع أو تتوقَّع رغبة الآخرين … وإذا كان، على نحو مشابه، مكُّوك النَّسيج سيحُوك خيوط الملابس من غير أيادٍ تغزلُه، وإذا كانت ريشة العازف ستضرب أوتار القيثارة بنفسها، فلا حاجة لربِّ العمل إذن بالخدم أو السادة بالعبيد.

جميعنا يُوافق أرسطو في ملاحظته حول حدوث انخفاضٍ فوري في العمالة حين يجد رب العمل وسيلةً آليةً لإنجاز العمل الذي كان يُنجزه العامل البشري سابقًا. والمُشكلة هنا هي ما إذا كانت الآثار الناتجة عن ذلك التَّحوُّل؛ «آثار التَّعويض»، والتي يميل إلى زيادة العمالة، ستُعوِّض حقًّا ذاك الانخفاض الحاصل أم لا. سيقول المتفائلون: نعم سيُعوِّض ذاك الانخفاض، وفي خضمِّ الجدال الحالي، ستراهم يُشيرون إلى جميع الوظائف الجديدة التي ظهرت بعد الثورات الصناعية السابقة. أما المُتشائمون فسيقولون: لا لن يحدث هذا، وسيُجادلونك بأن الآلات هي التي ستتولى إنجاز جميع تلك «الوظائف الجديدة» أيضًا. عندما تحل الآلات مكاننا في الأعمال البدنيَّة الجُهد، يمكن أن نتَّجه إلى الاشتغال بالأعمال الذهنية. لكن ماذا إذا حلَّت الآلات مكاننا أيضًا في إنجاز كل ما يتطلَّب مجهودًا ذهنيًّا، فما الذي بقي لنا؟

صوَّر ماكس تيجمارك هذا الجدال في كتابه «الحياة ٣٫٠» كحوارٍ بين حصانَين حول ظهور مُحرِّك الاحتراق الدَّاخلي في عام ١٩٠٠. تنبأ أحد الحصانين بـ «وظائف جديدةٍ للأحصنة. ... هذا هو دأب الحياة دائمًا، كما هو الحال عندما اختُرعت العجلة والمحراث». ولكن ما حدث للأسف أنَّ «الوظيفة الجديدة» لمعظم الأحصنة كانت أن يُصنع من لحمها طعام للحيوانات المنزلية الأليفة.

ظلَّ هذا الجدال مُتَّقدًا لآلاف السنين؛ لأنَّ هناك تأثيرات في كلا الاتجاهَين. والنتيجة الحقيقية تتوقَّف على كون أيٍّ تلك التأثيرات أهم لنا. ومثال ذلك، ما حدث لعُمال طلاء المنازل عندما تطوَّرت التقنية. ولتسهيل تصوُّر الأمر، سأستخدِم عرض فُرشاة الطلاء لأوضِّح درجة الأتمتة:

- إذا كانت الفُرشاة بعرض شعرةٍ واحدة (حوالي عُشر ملِّيمتر)، فسيستغرق طلاء منزلٍ واحدٍ حياة آلاف البشر؛ ومن ثم لا أحد سيعمل في طلاء المنازل.
- إذا كان لدَينا فرشاة بعرض ١ مليمتر، فرُبَّما وجدنا بعض الجداريات الصَّغيرة مطليَّة في القصر الملكي على يد حفنةٍ من الرسَّامين. وإذا كان لدَينا فرشاة بعرض ١ سنتيمتر، فسنجد الطبقة النَّبيلة كلها ستحذو حذو القصر الملكي.
- ما إن نحصل على فرشاةٍ بعرض ١٠ سنتيمترات (٤ بوصات)، فسنُفكِّر في الأمر بطريقةٍ عمليةٍ، وسنجد أنَّ معظم أصحاب المنازل سيطلُون بيوتهم من الداخل والخارج، رُغم أنهم لن يُكرِّروا طلاء منازلهم في وقتٍ قصير، وسيجد الآلاف من عمال طلاء المنازل عملًا لهم.

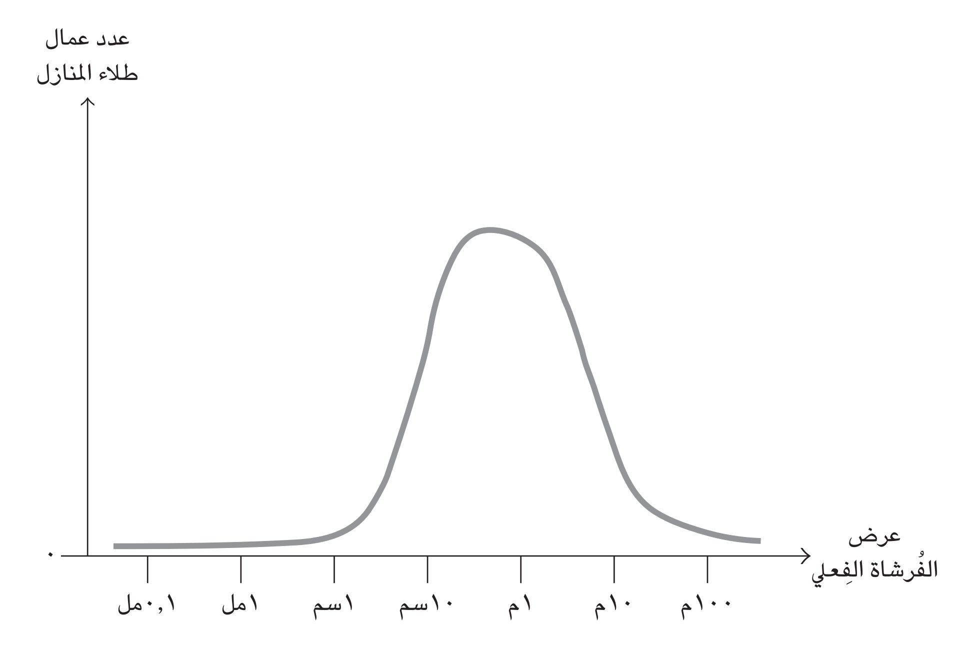
- عندما نحصُل على الفُرشات الأسطوانية ورشاشات الطِّلاء (والتي تُعادل فرشاةً بعرض مترٍ واحد تقريبًا)، فإن التكلفة ستنخفض انخفاضًا كبيرًا، لكنَّ السُّوق حينها قد يبدأ في التَّشبُّع ويقلُّ الطَّلب، فيبدأ عدد عُمال طلاء المنازل بالانخفاض بعض الشيء.
- عندما يدير شخص واحد فريقًا من مائة روبوت لطلاء المنازل (بإنتاجيةٍ تُعادل فرشاة بعرض ١٠٠ متر) فإنَّ منازل بأكملها يمكن أن تُطلى في ساعةٍ واحدةٍ، ولكن لن يكون هُناك سوى عددٍ قليل جدًّا من عمال الطلاء البشريِّين الذين يعملون في هذه المهنة.

بالتالي، فإن التأثير «المباشر» للتطوُّر التقني يعمل في كلا الاتجاهين؛ في بادئ الأمر، مع زيادة الإنتاجية، يمكن أن تزيد التقنية من العمالة عبر تخفيض تكلفة العمل وبالتالي يزداد الطلب عليه، ولكن لاحقًا، كلَّما تطوَّرت التقنية أكثر، قلَّ عدد العمالة البشرية المطلوبة أكثر فأكثر. والشكل ٤-٢ يُوضِّح تلك التَّطورات.[20]

تنتج العديد من التِّقنيات مُنحنيات مُشابهة. وإذا كُنا، في أي قطاعٍ من القطاعات الاقتصادية، على يسار لُمنحنًى، فإنَّ هذا يعني أنَّ تطوُّر التقنية يزيد من الوظائف في هذا القطاع. والأمثلة في واقعنا المُعاصر قد تشمل مهام مثل إزالة رسوم الجُدران، والتنظيف البيئي، وتفتيش حاويات الشَّحن، وبناء المنازل في البُلدان الأقل تطورًا، والتي جميعها قد تُصبح ذات جدوى اقتصادية أكبر إذا ما أُنجزت بمساعدة الروبوتات لنا. أما إذا كُنا في الجانب الأيمن من المُنحنى، فإنَّ زيادة الأتمتة ستُقلِّل من العمالة. فمثلًا، ليس من الصعب التوقُّع أنَّ مهنة عامل المصعد ستستمر في التَّقلُّص حتى تختفي. على المدى البعيد، يحسُن بنا التَّوقُّع أنَّ معظم الصناعات ستُدفع دفعًا إلى أقصى يمين المنحنى. في وقتٍ قريبٍ، نشر عالما الاقتصاد ديفيد أوتار وآنا سالومنز مقالًا مبنيًّا على دراسةٍ متأنِّيةٍ في مجال الاقتصاد الإحصائي يُقِرُّ بأنَّ «على مدار الأربعين سنةً الماضية، انخفضت الوظائف في جميع الصناعات التي أدخلت الحلول التقنية لزيادة إنتاجيتها».[21]

ولكن ماذا عن «آثار التَّعويض» التي وصفها الاقتصاديون المُتفائلون؟

- بعض الناس سيعملون في صناعة روبوتات الطلاء. كم عددهم؟ أقل «بكثيرٍ» من عدد عُمال الطلاء الذين حلت محلهم الروبوتات؛ وإلا فإن تكلفة طلاء



شكل ٤-٢: رسم بياني تصوري للعمالة في مجال طلاء المنازل مع تطوُّر تقنيات الطلاء.

المنازل سترتفع في حالة استخدام الروبوتات (ولن تقل)، وحينها لا أحد سيشتري الروبوتات.

- سيُصبح طلاء المنازل أقل تكلفةً بعض الشيء، وحينها سيتَّجه الناس إلى طلاء منازلهم مراتٍ أكثر قليلًا.
- وأخيرًا، لأنَّنا ندفع أقل في طلاء المنازل، فسيكون لدَينا مالٌ أكثر لنصرفه على شراء أشياء أخرى، وهكذا نزيد فُرص العمل في مجالاتٍ أخرى.

حاول الاقتصاديون قياس حجم تلك الآثار في العديد من الصناعات التي تشهد زيادةً في الأتمتة، لكنَّ النتائج غير نهائية بوجهٍ عام.

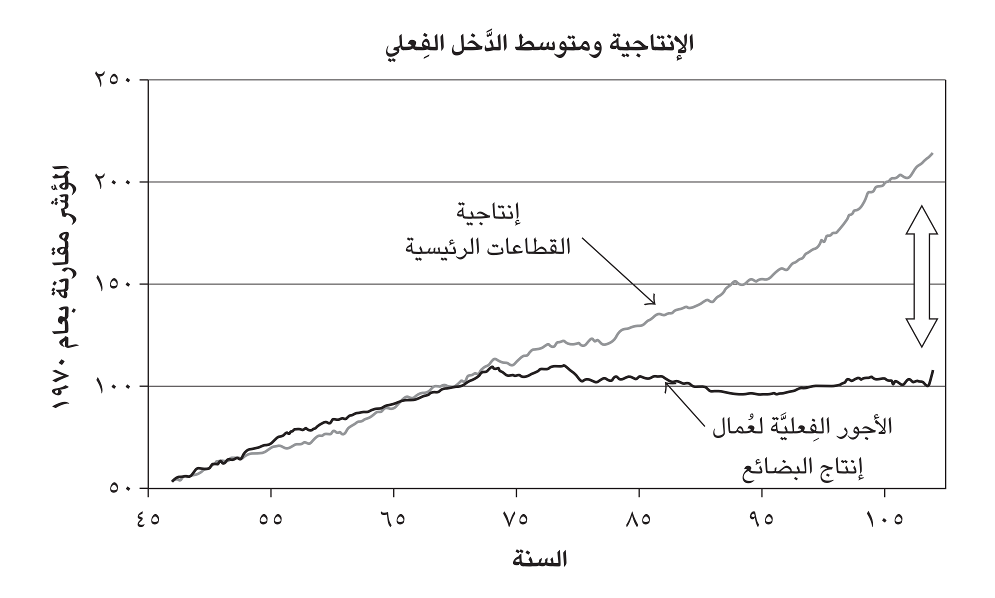
عبر التاريخ، كان الاتجاه السائد لدى معظم علماء الاقتصاد الذين ناقشوا هذه القضية، هو النظر إليها باستخدام «الصورة الكلية»: الأتمتة تزيد الإنتاجية، إذن، بنظرةٍ

عامة، سيكون البشر أفضل حالًا من ناحية أنَّنا سنستمتع بالمزيد من البضائع والخدمات بنفس القدر من العمل.

للأسف، النظرية الاقتصادية لا تتنبأ بأن جميع البشر سيكونون أفضل حالًا نتيجةً للأتمتة. الأتمتة بوجه عام تزيد من حصَّة الدَّخل التي تصبُّ في رأس المال (أي أصحاب آليي طلاء المنازل) وتُنقص من حصَّة الدَّخل التي تصبُّ في العمالة (أي عمال طلاء المنازل السابقين). يرى عالما الاقتصاد إريك براينجولفسن وأندرو ماكافي في كتابهما «عصر الآلة الثاني»، أنَّ هذا الأمر يحدث بالفعل منذ عدة عقود. إن بيانات الولايات المتحدة الأمريكية مُوضَّحة في الشكل ٤-٣، وتُشير إلى أنَّ الأجور والإنتاجية في الفترة ما بين ١٩٤٧ و١٩٧٣، ارتفعتا معًا، ولكن بعد عام ١٩٧٣، ثبتت الأجور بينما أخذت الإنتاجية تزداد حتى تضاعفت. ويُطلِق الكاتبان على هذا اسم «الانفصال العظيم». وهناك المزيد من علماء الاقتصاد البارزين الآخرين الذين حذَّروا من هذا الخطر، من بينهم علماء حصلوا على جائزة نوبل وهم روبرت شيلر ومايك سبينس وبول كروجمان؛ وكلاوس شواب رئيس المنتدى الاقتصادي العالمي؛ ولاري سامرز كبير الخبراء الاقتصاديين بالبنك الدولي ووزير المالية في عهد الرئيس الأمريكي بيل كلينتون.

عادةً ما كان يُشير هؤلاء الذين يختلفون مع فكرة البطالة التِّقنية إلى وظائف مثل وظيفة الصراف داخل أروقة المصارف، والذي يُمكن للصَّراف الآلي أن يُنجز عمله جزئيًّا، ووظيفة صراف متاجر التجزئة الذي صار يُنجز عمله بسرعةٍ بفضل الأرقام التَّسلسُليَّة (أكواد الباركود) وعلامات رقاقات الراديو اللاسلكية (آر إف آي دي) الملصوقة على البضائع والمنتجات. فهُم عادة ما يدَّعون أنَّ هذه الوظائف تزدهر «بفضل» التطوُّر التِّقني. والحقيقة أنَّ الواقع يُصدِّق هذا الكلام؛ فأعداد الصَّرافين في الولايات المتحدة الأمريكية قد تضاعفت تقريبًا في الفترة ما بين عامي ١٩٧٠ و٢٠١٠، ومع ذلك فمن المُهمِّ أن نعرف أنَّ في نفس الفترة ازداد عدد السُّكان بنسبة ٥٠ بالمائة، وازدهر القطاع المالي بنسبة تزيد على ٤٠٠ بالمائة،[22] لذلك من الصَّعب أن نُرجع كل الفضل، أو أي فضلٍ مُطلقًا، في هذه الزيادة في الوظائف إلى الصَّرافات الآلية. وللأسف، في الفترة ما بين ٢٠١٠ و٢٠١٦، فقد نحو ١٠٠ ألف صرافٍ وظائفهم، ومن المُتوقَّع طبقًا لمكتب إحصاءات العمل الأمريكي أن يفقد ٤٠ ألفًا آخرون وظائفهم بحلول عام ٢٠٢٦: «الصَّيرفة الإلكترونية والأتمتة يُتوقَّع أن تستمرَّا في إنجاز المزيد والمزيد من المهام التي كان الصَّرافون عادةً ما ينجزونها».[23] والبيانات المُتاحة بخصوص صراف متاجر التجزئة لا تُبشِّر بخيرٍ؛ فقد

انخفض المُعدَّل الفردي ٥ بالمائة في الفترة ما بين عامي ١٩٩٧ و٢٠١٥، ويُخبرنا مكتب إحصاءات العمل أنَّ «التَّطورات التِّقنية مثل منصَّات الدَّفع الذاتي في متاجر التجزئة وازدياد حركة التَّسوق الإلكتروني، ستستمرُّ في تقليل الحاجة لعمل الصرَّافين في المتاجر.» يبدو لنا أنَّ كلا القطاعين قد بدأ رحلته على منحى الهبوط، والأمر ينطبق على جميع المهن المُنخفضة المهارة تقريبًا، التي تتطلَّب العمل جنبًا إلى جنبٍ مع الآلات.



شكل ٤-٣: بيانات الإنتاج الاقتصادي ومتوسط الأجور الفعلي في الولايات المتحدة الأمريكية منذ عام ١٩٤٧. (البيانات مأخوذة من مكتب إحصاءات العمل الأمريكي.)

إذن، ما هي الوظائف التي على وشك الاختفاء مع وصول تقنياتٍ جديدةٍ قائمة على الذكاء الاصطناعي؟ المثال الرئيسي لهذا النَّوع من الوظائف والذي يُضرب دائمًا في وسائل الإعلام هو وظيفة القيادة. هناك ما يقرُب من ٣٫٥ ملايين سائق شاحنةٍ في الولايات المُتَّحدة الأمريكية، والعديد من تلك الوظائف سيكون لا محالة عُرضة للأتمتة. إن شركة أمازون، وغيرها من الشركات الأخرى، تستعمل حاليًّا شاحنات ذاتية القيادة لنقل البضائع على

الطرق السريعة ما بين الولايات، ولكن بوجود سائقين بشريِّين احتياطيِّين.[24] ويبدو من المُحتمل جدًّا عما قريب أن يُصبح الجزء الأطول من رحلة النَّقل على الطرق السريعة مُؤتمتًا كُليًّا، بينما سيتولَّى السائق البشري في الوقت الحالي القيادة داخل المدينة وعملية استلام البضاعة وتسليمها. ونتيجةً لهذه التَّطورات المُتوقَّعة، فإن عددًا قليلًا جدًّا من الشباب لدَيه اهتمام بقيادة الشاحنات كمهنةٍ؛ ومن المُثير للسُّخرية أنَّ هناك نقصًا حادًّا في سائقي الشاحنات حاليًّا في الولايات المتحدة الأمريكية، مما يدفع فجر الأتمتة إلى بزوغٍ مُعجَّل.

لم تسلم الوظائف الإدارية أيضًا من خطر الأتمتة. على سبيل المثال، يتوقَّع مكتب إحصاءات العمل الأمريكي أن تنخفض نسبة العمالة في وظيفة وكلاء التأمين بنسبة ١٣ بالمائة في الفترة ما بين عامي ٢٠١٦ و٢٠٢٦: «إن برمجيات التأمين الآلية تُتيح للعاملين أن يُنجزوا استماراتهم على نحوٍ أسرع من ذي قبل، ممَّا يُقلِّل الحاجة على نحوٍ كبير إلى وكلاء التأمين.» وإذا تطورت التقنيات اللُّغوية كما هو مُتوقَّع، فالعديد من وظائف خدمة العملاء والمبيعات ستكون عُرضةً أيضًا للأتمتة، كما ينطبق هذا الكلام أيضًا على الوظائف القانونية. (في عام ٢٠١٨، تفوَّق برنامج ذكاء اصطناعي على أساتذة قانونٍ مُتمرِّسين في تحليل اتفاقات عدم إفصاحٍ نموذجية، وأنهى المُهمَّة أسرع بـ ٢٠٠ مرة.)[25] حتى الجوانب النَّمطيَّة في مجال برمجة الكمبيوتر، التي من النوع الذي يجري تعهيده عادة اليوم، هي الأخرى عُرضة للأتمتة. إن تقريبًا أي عملٍ يُمكن تعهيده هو بلا شكٍّ مُرشَّح جيد للأتمتة، وهذا لأنَّ عملية التعهيد ما هي إلا تقسيم العمل إلى مهامَّ صغيرةٍ يُمكن توزيعها والعمل عليها خارج سياق المشروع الرئيسي. وتنتج صناعة «أتمتة العمليات باستخدام البرامج الآلية» أدوات برمجية تُحقق نفس هذا الشيء في المهام الإدارية المنجزة عبر الإنترنت.

ومع تقدُّم الذكاء الاصطناعي، بالطبع من المُمكن (بل من الوارد جدًّا) أن خلال العقود القليلة القادمة، ستُنجز جميع الأعمال النَّمطيَّة؛ البدنية منها والذهنية، بتكلفةً أقل على يد الآلات. ولأنَّنا لم نعُد نصطاد ونجمع الثمار في جماعاتٍ كما اعتدنا منذ آلاف السنين، فإن مُجتمعاتنا استخدمت معظم الناس ليكونوا مثل الروبوتات لأداء مهامَّ يدويةٍ وذهنية مُتكرِّرة، لذلك ربما ليس من المُستغرب أن تحل الروبوتات مكاننا قريبًا في تلك الأدوار. وعندما يحدث هذا، ستنخفض أجور أولئك الذين لا يستطيعون المُنافسة على الوظائف المُتبقِّية ذات المهارات العالية إلى ما تحت مُستوى خط الفقر. يصُوغ لاري سامرز هذا الأمر قائلًا: «قد يصل الأمر، إذا وضعنا احتمالات وجود بدائل أمام أرباب

الأعمال لاستبدال العمالة بالروبوتات، إلى أنَّ بعض قطاعات الوظائف لن تستطيع حتى أن تكسب قُوت يومها لتعيش حدَّ الكفاف».[26] وهذا بالضَّبط ما حدث للأحصنة؛ فقد صارت وسائل المواصلات الميكانيكية أقلَّ تكلفةً إذا ما قُورنت بتكلفة رعاية أحد الأحصنة، لذلك أصبحت الأحصنة طعامًا للقطط والكلاب. وعندما يُواجَه البشر بالمقابل الاجتماعي والاقتصادي لأن يكونوا طعامًا للحيوانات الأليفة، فإنهم سيكونون ساخطين أشدَّ السخط على حكوماتهم.

ونظرًا لاحتمالية مواجهتها لسخط مواطنيها، فإن الحكومات حول العالم بدأت بالفعل في الانتباه إلى هذه المُشكلة. ومعظمها قد أدرك الآن أن فكرة إعادة تأهيل الجميع ليكونوا علماء بيانات أو مهندسي روبوتات لن تُجدي نفعًا؛ فالعالم قد يحتاج إلى خمسة أو عشرة ملايين من هؤلاء، لا جميع هؤلاء المليار موظَّفٍ الذين على وشك خسارة وظائفهم. إن مجال علوم البيانات ما هو إلا قارب نجاةٍ صغير لن يحمل جميع رُكاب الباخرة العملاقة الغارقة.[27]

يُعدُّ البعض «خُططًا انتقالية»، ولكن السؤال هو: انتقالية إلى ماذا؟ نحن نحتاج أن يكون لدَينا وجهة واضحة لنضع خُطَّة انتقالية؛ أي نحتاج صورة واضحةً لاقتصادٍ مستقبلي مقبول تُنجز فيه الآلات مُعظم ما نُسمِّيه اليوم عملًا.

أحد صور الاقتصاد المُستقبلي التي تظهر على الساحة في تسارُعٍ هي حيث يكون هناك طائفة أقل كثيرًا من الناس يعملون في وظائف لأنَّ العمل ليس شيئًا ضروريًّا. وقد تخيَّل كينز هذه الصورة المُستقبلية في مقاله «الخيارات الاقتصادية لأحفادنا». ووصف موجة البطالة العارمة التي ابتُليت بها بريطانيا في عام ١٩٣٠ على أنَّها «موجة مؤقَّتة من عدم التَّوازن» تسبَّبت فيها «زيادة الكفاءة التقنية» التي حدثت «بوتيرةٍ أسرع ممَّا يُمكننا التعامل مع مشكلة استيعاب اليد العاملة». لكنَّه، رُغم ذلك، لم يتخيَّل أن على المدى البعيد، بعد قرنٍ من الزمان مليء بالتَّطورات التقنية، ستكون هناك عودة لاستيعاب جميع الأيدي العاملة في سوق العمل، فقال:

> وهكذا، ولأول مرةٍ منذ أن خُلق الإنسان، سيُواجه مشكلته الحقيقية والدائمة، وهي: كيف سيستفيد بعد تحرُّره من وطأة المشاغل الاقتصادية المُلحَّة، وكيف سيستمتع بالراحة التي سيكون العلمُ وأموال الفوائد المُركَّبة قد وفَّراها له ليعيش حياةً رشيدةً ومتناغمةً وينعم بالعافية.

مثل هذا المُستقبل يتطلَّب تغييرًا جذريًّا في نظامنا الاقتصادي؛ لأنَّ في الكثير من دول العالم، أولئك الذين لا يعملون يُواجهون الفقر أو العوز. ولذلك، ستجد أنصار رؤية كينز المُعاصرون عادةً ما يدعمون توفير شكلٍ ما من أشكال «الدَّخل الأساسي العام». إن هذا الدخل، المُموَّل من ضرائب القيمة المُضافة أو ضرائب عائد رأس المال، سيوفِّر مستوًى معيشيًّا مقبولًا لجميع البالغين بصرف النَّظر عن ظروفهم. أما الذين يطمحون للعيش في مُستوًى أفضل، فيمكنهم العمل من غير أن يفقدوا هذا الدخل الأساسي، وأولئك الراضُون بمُستوى معيشتهم، يمكنهم أن يفعلوا ما يحلو لهم في وقتهم. ورُبَّما من المُدهش أن فكرة الدَّخل الأساسي العام مدعومة من جميع الأطياف السياسية؛ بدايةً من معهد آدم سميث[28] وحتى حزب الخضر.[29]

بالنسبة إلى البعض، الدخل الأساسي العام يُمثِّل نُسخةً أرضيةً من الجنَّة.[30] بينما تراه طائفة أخرى من الناس أنَّه يعني اعترافًا بالفشل؛ فهم يرون أنَّ معظم الناس بذلك لن يملكوا أي قيمةٍ اقتصاديةٍ ليُساهموا بها في المُجتمع؛ فهم سيُطعمون ويُسكَّنُون في المنازل (غالبًا على يد الآلات)، وفيما عدا ذلك، سيُتركُون إلى إرادتهم الحُرَّة. والحقيقية، كما هي دائمًا، في مكانٍ ما بين الرأيين وتعتمد اعتمادًا كبيرًا على رؤية المرء لطبيعة النَّفس الإنسانية. لقد فرَّق كينز في مقاله بين أولئك الذين يُكافحون ويسعون وبين أولئك الذين يتمتَّعون؛ أولئك «الطَّموحين» الذين يسعون بُكلِّ جُهدهم وراء متع مستقبلية، وبين أولئك «المُبتهجين» الذين «يستطيعون الاستمتاع المباشر بالأشياء». ومُقترح الدَّخل الأساسي العام يفترض أنَّ السَّواد الأعظم من الناس سيكونون من زُمرة الأشخاص المُبتهجين.

يرى كينز أنَّ السَّعي هو أحد «عادات وغرائز البشر والتي قد غُرست بداخلهم جيلًا بعد الآخر منذ أمدٍ بعيد» وليست «قيمةً حقيقةً من قيم الحياة». كما يتنبأ أنَّ هذه الغريزة ستندثر شيئًا فشيئًا حتى تختفي. وخلافًا لوجهة النَّظر هذه، قد يرى أحدهم أنَّ السَّعي هو جوهر كون الفرد إنسانًا حقيقيًّا. وبدلًا من رؤية السعي والاستمتاع على أنَّهُما شيئان مُنفصلان لا يلتقيان، فإنَّهُما غالبًا ما يُلازم أحدهما الآخر؛ فالمُتعة الحقيقية والإحساس الدائم بروعة الإنجاز يتأتَّيان من وجود غايةٍ ما وتحقيقها (أو على الأقل مُحاولة تحقيقها)، غالبًا في مواجهة الصِّعاب والعقبات، وليس من الاستهلاك السَّلبي للمُتع المُباشرة؛ فهناك فرق بين تسلُّق جبل إيفرست وبين أن تُنقل إلى قمَّته بطائرةٍ مروحية.

والعلاقة بين السعي والاستمتاع هي موضوع محوري لفَهمنا كيفية صياغة مُستقبلٍ جيدٍ. ربما ستتساءل الأجيال القادمة عن سبب قلقنا حول ذلك الشيء العقيم الذي بلا

فائدة الذي يُسمَّى «العمل». وتحسُّبًا لأن يكون هذا التَّغيُّر في الرؤى سيُحدُث على نحوٍ بطيءٍ، دعونا إذن نتفكَّر في التبِعات الاقتصادية لوجهة النَّظر التي ترى أنَّ أحوال مُعظم الناس ستكون جيدةً إذا كان لهم دور نافع ليقوموا به، حتى لو كانت مُعظم البضائع والخدمات ستُنتج على يد الروبوتات بإشراف بشري يكاد لا يُذكَر. حينها، لا محالة أنَّ معظم الناس سينخرطُون في تقديم الخدمات التفاعلية التي يُمكن للبشر فقط تقديمها، أو بالأحرى، تلك التي «نُفضِّل» أن يُقدِّمها البشر. هذا يعني أننا إذا كُنا من الآن فصاعدًا لن نستطيع أن نُساهم بأي عملٍ بدنيٍّ أو ذهنيٍّ روتيني، فأقل القليل أن نُساهم بإنسانيَّتنا. وحينها سنحتاج أن نبرع في أن نكون بشرًا.[31]

والمهن الحالية التي من هذا النَّوع تشمل المُعالِجين النَّفسيين، ومُوجهي المديرين التنفيذيين والمُعلِّمين والمُستشارين والمساعدين وجُلساء الأطفال وكبار السِّن. وعبارة «مهن الرِّعاية» غالبًا ما تُستخدَم في هذا السِّياق، لكنِّي أراها عبارةً مُضلِّلة؛ فتلك العبارة لها بالتأكيد وقْعٌ إيجابي في أذُن مُقدِّمي الرِّعاية، بينما لها أثر سلبيٌّ يُخبرنا عن مدى اعتمادية وعجز مُتلقِّي تلك الرعاية. لكن لنعُد إلى مقال كينز مرةً أخرى ونتفكَّر في تلك المُلاحظة:

> إن الذين استطاعوا البقاء على قيد الحياة وصقل مهاراتهم حتى تصل إلى حدِّ الكمال في فنِّ الحياة، ولا يشترون بأنفُسهم سُبُل الحياة الوضيعة هم الذين سينعمُون بالحياة الرَّغدة حين تأتي.

جميعنا نحتاج إلى مساعدة في تعلُّم «فن الحياة». هذه ليست مسألة اعتمادية، بل مسألة نُمُو. إن القدرة على إلهام الآخرين وإكسابهم حس التَّذوق والإبداع — في الفن أو الموسيقى أو الأدب أو المُحادثة مع الغير أو البستنة أو الفنون المعمارية أو الطعام أو الشَّراب أو ألعاب الفيديو — سنحتاج إليها على الأرجح أكثر من أي وقتٍ مضى.

المسألة التالية هي توزيع الدَّخل. في أغلب البلدان، هذا الأمر ينحرف إلى طريقٍ خاطئ مُنذ عدة عُقودٍ. إنها مسألة معقَّدة، ولكنَّ هناك شيئًا واحدًا واضحًا كالشَّمس؛ وهو أن الدَّخل المُرتفع والحالة الاجتماعية العالية غالبًا ما يتأتَّيان من تقديم قيمةٍ مضافة عالية. ولنضرب مثالًا؛ مهنة مجال رعاية الأطفال تُربط بالدَّخل المُنخفض والحالة الاجتماعية المُتدنِّية. وهذا راجع في بعضٍ منه كنتيجةٍ لجهلنا بأُسُس تلك المهنة وكيفية أدائها. بعض المُشتغِلين بهذا يُؤدُّونها غريزيًّا على نحوٍ جيد، لكنَّ الأغلبية ليسوا كذلك. قارن هذا مثلًا بمهنة جراحة العظام. ببساطةٍ، لن نذهب نحن إلى مُراهقٍ ملول يحتاج إلى المال ثمَّ نختاره

للعمل كجرَّاح عظامٍ لقاء خمسة دولاراتٍ في الساعة إلى جانب السَّماح له بحشو معدته بما يُريد من ثلاجة المنزل. لقد استثمرنا قُرونًا من البحث لمعرفة جسد الإنسان وكيفية علاج أجزائه حين يحدث بها عطب، وجرَّاح العظام عليه أن يخضع لسنواتٍ من التدريب ليحصُل على كل هذه المعرفة والمهارات المطلوبة لتطبيقها. ولهذا، فإنَّ جرَّاحي العظام يحصلون على دخلٍ مُرتفعٍ ويتمتَّعون بمكانةٍ اجتماعيةٍ راقية. وهم لا يحصلون على دخلٍ مُرتفع فقط لأنَّ لديهم الكثير من المعرفة ويخضعون للكثير من التدريب، بل أيضًا لأنَّ جميع تلك المعرفة والتدريب تُؤتي ثمارها. فهي تُمكِّنُهم من المُساهمة بقيمةٍ كبيرة في حياة الآخرين، خصوصًا ذوي العظام المكسورة.

لسوء الحظ، معرفتنا العلمية بآلية عمل الدِّماغ ضعيفة على نحوٍ صادم، ومعرفتنا العلمية بأمورٍ مثل السَّعادة والإنجاز أشدُّ ضعفًا. نحن ببساطةٍ لا نعرف كيف نُضيف قيمةً في حياة بعضنا لبعض على نحوٍ مُطَّردٍ وقابل للتوقُّع. صحيح أننا حقَّقنا نجاحًا مقبولًا في فهم بعض الاضطرابات النفسية، لكنَّنا ما نزال نُحارب منذ فترةٍ طويلة في معركةٍ تعليميةٍ حول شيءٍ بسيطٍ كتعليم القراءة للأطفال.[32] إننا نحتاج إلى إعادة النظر جذريًّا في نظامنا التعليمي ومؤسَّساتنا العلميَّة لنضع جُلَّ تركيزنا على الإنسان بدلًا من التركيز على العالم المادي. (يرى جوزيف آون، رئيس جامعة نورث إيسترن، أنَّ الجامعات يجب أن تُدرِّس وتدرُس «علم الطبيعة البشرية».)[33] قد يبدو من الغريب القول إن السعادة يجب أن تكون علمًا هندسيًّا، لكن يبدو أنَّه لا مناص من هذا. إن مثل هذا العلم سيُبنى على العلوم الأساسية — أي فهم أفضل لآلية عمل الدِّماغ البشري على المستويين المعرفي والعاطفي — وسيُؤهِّل العديد من المُمارسين في مجالاتٍ تتنوَّع ما بين مُهندسي الحياة، أولئك الذين سيُساعدون الأفراد على التَّخطيط لمسارات حياتهم بأكملها؛ والخبراء المهنيِّين في مجالاتٍ كمجال تعزيز غريزة الفُضُول وحب الاستطلاع، والتَّكيُّف الشخصي والصمود أمام الصُّعوبات. وإذا كانت تلك المهن ستُبنى على أُسُسٍ علميةٍ سليمة، فعليها أن تكون منطقيةً وعقلانيةً كمهنة المُهندس الذي يُصمِّم جسرًا أو جراح العظام في وقتنا الحاضر.

إعادة النظر في مؤسستنا التعليمية والبحثية، لتوفير تلك العلوم الأساسية ولتحويلها إلى برامج تدريبية وتخريج أفراد مؤهَّلين، ستستغرق عُقودًا من الزمان، لذلك أظنُّها فكرةً جيدةً أن نبدأ الآن، ويا لها من حسرةٍ أنَّنا لم نبدأ منذ زمنٍ بعيد. والنتيجة النهائية

(إن نجح الأمر) ستكون عالمًا يستحقُّ أن نحيا فيه. أما بدون عملية إعادة النظر هذه، فإننا نُخاطر بمستوًى غير مُحتمل من الاضطراب الاجتماعي والاقتصادي.

(٤) الاستيلاء على أدوار أخرى للبشر

علينا أن نُفكِّر جيدًا قبل أن نسمح للآلات بأن تضطلع بأدوار تشمل خدماتٍ تفاعلية بين الأفراد. وإذا جاز القول إن إنسانيتنا هي نقطة قوتنا الرئيسية في التعامُل مع غيرنا من البشر، حينها سيبدو صُنعُ آلاتٍ تُحاكي البشر فكرةً سيئةً. لُحسن حظِّنا، نحن البشر لدَينا ميزة واضحة نتفوَّق بها على الآلات في أمر معرفة ما يشعر به غيرنا من البشر وكيف سيتصرَّفون. إن جميع أفراد الجنس البشري تقريبًا يعرفون ماهيَّة شعور أن يضرب المرء إبهامه بمطرقة، أو يُحبَّ حبًّا غير مُتبادل.

وعدم استغلال هذه الميزة البشرية الفطرية وإبطالها، هو عيب بشري فطري؛ فنحن ميالون إلى أن نُخدع بالمظاهر، وخصوصًا المظاهر البشرية. وقد حذَّر آلان تورينج من صُنع روبوتات تُشبه البشر، فقال:[34]

> أرجو بل وأؤمن أنَّنا بلا شكٍّ لن نبذُل جهدًا في صُنع آلاتٍ تحمل أكثر صفات البشر غير الفكرية تميُّزًا مثل أن يكون لها أجساد كأجساد البشر؛ فأرى من وجهة نظري أن مثل هذا الصَّنيع إنما هو صنيع عقيم ونتائجه ستكون لها نفس الجودة الرديئة التي لصُنع ورودٍ صناعية.

للأسف، ذهب تحذير آلان أدراج الرياح ولم نُعره أي اهتمام. فالعديد من المجموعات البحثية قد أنتجت روبوتات على هيئةٍ بشريةٍ واقعيةٍ على نحوٍ مُخيف، كأنَّهم ينبضون بالحياة كما هو مُوضَّح في الشكل ٤-٤.

إذا نظرنا إلى الروبوتات كأدواتٍ بحثيةٍ، فقد نستخلص منهم رُؤًى حول كيفية تفسير البشر لسلوك الروبوتات وتواصلهم. أما إذا نظرنا إليهم كنماذج أولية لُمنتجاتٍ تجارية مُستقبلية، فإنهم سيُمثِّلون نوعًا من التضليل والكذب. فهُم يتجنَّبون وعينا الُمدرك ويُخاطبون عواطفنا مُباشرةً، وُربَّما يُقنعُوننا بأنَّهم قد وُهبوا ذكاءً حقيقيًّا. تخيَّل مثلًا مدى سُهولة أن تُغلق وتُعيد تشغيل روبوت على شكل صندوقٍ رمادي جاثم لأنَّ به مشكلة ما (حتى ولو كان يملأ الدنيا صياحًا ويُخبُرك أنَّه لا يُريد أن يُطفأ)، وما هي صعوبة فعل نفس الشيء مع روبوتات مثل «جيا جيا» أو «جيمينويد دي كيه». تخيَّل أيضًا كم

سيكون مُربكًا وربما يُسبِّب اضطراباتٍ نفسيةً للأطفال والرُّضَّع إذا وضعوا تحت رعاية روبوتات تبدو مثل البشر، مثل آبائهم، لكنَّهم ليسوا كذلك؛ ويُظهرون العطف والرِّعاية، مثل آبائهم، لكنَّهم في الحقيقة ليس لديهم مشاعر أصلًا.





شكل ٤-٤: (على اليمين) «جيا جيا»، الروبوت الذي صُنع في جامعة العلوم والتِّقنية الصينية. (على اليسار) «جيمينويد دي كيه»، الروبوت الذي من تصميم هيروشي إشيجورو من جامعة أوساكا اليابانية، والذي صُمِّم لمُحاكاة وجه هينريك شيرفا من جامعة آلبورج الدانمركية.

لا فائدة حقيقية تُرجى من صُنع روبوتات على هيئةٍ بشريةٍ إلا فيما عدا القُدرة الأساسية على توصيل المعلومات غير اللفظية عبر تعبيرات الوجه وحركات تقاسيمه؛ تلك التي استطاعت حتى الشخصية الكارتونية «بجز بني» أداءها بسهولةٍ ويُسر. وهناك أيضًا أسباب وجيهه وعمليَّة تدفعُنا ألا نضع الروبوتات في قالبٍ بشري؛ مثلًا، هيئتنا نحن البشر الواقفة على قدمَين أقل ثباتًا إذا ما قُورنت بالمشي على أربع. إن القطط والكلاب والأحصنة تندمج مع حياتنا البشرية على نحوٍ جيدٍ وهيئتها البدنيَّة دليل واضح جدًّا على طريقة تصرُّفها المتوقعة. (تخيَّل أنَّ حصانًا بدأ يتصرَّف فجأة ككلب!) وهذا الأمر يجب أن ينطبق على الروبوتات أيضًا. رُبَّما هيئة لها أربع أرجلٍ وذراعان وتركيب جسدي على هيئة كائن القنطور الأسطوري سيكون نموذجًا قياسيًّا مقبولًا. أما أن تُحاكي الروبوتات البشر في

جميع التَّفاصيل، فهو يُشبه صُنع سيارة فيراري سُرعتها القُصوى ٥ أميالٍ في الساعة، أو مُثلَّجاتٍ بطعم التُّوت، لكنَّها في الحقيقة مصنوعة من معجون شرائح الكبدة المصبوغ بلون البنجر الأحمر.

تلك الهيئة البشرية التي لبعض الروبوتات قد تسبَّبت بالفعل في بعض الارتباك السياسي والعاطفي. في الخامس والعشرين من أكتوبر ٢٠١٧، منحت المملكة العربية السُّعودية الجنسية السعودية للروبوت «صوفيا»؛ وهو روبوت على هيئةٍ بشريةٍ قد وُصف بأنَّه لا يعدو كونه «نظام دردشة لديه وجه»،[35] بل أسوأ.[36] ربما كانت هذه الواقعة حركة استعراضية في مجال العلاقات العامة، لكن أن يُصدر مقترح من لجنة الشئون القانونية بالبرلمان الأوروبي، لهُو إذن أمر جاد وخطير.[37] فهذا المُقترح كان يُوصي بالآتي:

> إعطاء صفةٍ قانونيةٍ خاصة للروبوتات على المدى الطويل، حتى يكون هناك على الأقل لأكثر الروبوتات تطورًا واستقلاليةً صفة الأشخاص الإلكترونيِّين حتى يكونوا مسئولين عن أي ضررٍ قد يتسبَّبون به.

بعبارةٍ أخرى، سيُصبح «الروبوت» مسئولًا أمام القانون عن أيِّ ضررٍ يُوقعُه، بصرف النَّظر عن صاحبه أو مُصنِّعه. وهذا يُوحي بأنَّ الروبوتات سيكون لهم أصول مالية وسيكونون عُرضة للعقوبات إن لم يلتزموا بالقوانين. إن هذا الكلام مُجرَّد هُراءٍ لا معنى له. مثلًا، إذا كُنا سنزُجُّ بأحد الروبوتات في السِّجن لعدم سداده المُستحقات المالية، فما الذي سيضيره إذا سُجن؟

بالإضافة إلى هذه المنزلة غير المُبرَّرة والغريبة التي تُرفع إليها الروبوتات، فإنَّ ثمة خطرًا مُحدقًا من زيادة استخدام الآلات في إصدار القرارات التي تمسُّ حياة الناس، لأنَّها ستُؤدِّي إلى الحطِّ من منزلة وكرامة البشر. وهذا الاحتمال قد صُوِّر بإتقانٍ في مشهدٍ من مشاهد أحد أفلام الخيال العلمي يُسمَّى «إليزيام»، حيثُ يقف المُمثِّل مات دايمن في شخصية ماكس ليترافع عن نفسه أمام «ضابط الإفراج المشروط» (انظر الشكل ٤-٥) ويشرح له لماذا يرى أن تمديد فترة عقوبته غير مُبرَّر. ولا حاجة للقول أنَّ سعي ماكس قد خاب، بل إن ضابط الإفراج المشروط قد وبَّخه لعدم إظهاره سلوكًا محترمًا.

يُمكن للمرء أن ينظر إلى هذا الاعتداء على الكرامة الإنسانية بطريقتَين. الطريقة الأولى هي المُباشرة؛ وهي أنَّ بإعطاء الآلات سُلطةً على البشر، فنحن نُنزل من أنفسنا كجنسٍ بشري إلى مرتبةٍ أقل ونفقد حق المشاركة في اتخاذ القرارات التي تمسُّ حياتنا.



شكل ٤-٥: ماكس (الذي يقوم بدوره المُمثِّل مات دايمن) وهو يُقابل ضابط الإفراج المشروط في فيلم «إليزيام».

(وإعطاء الآلات السُّلطة لقتل البشر، كما ناقشنا في نقطةٍ سابقةٍ في هذا الفصل، هو مثال أكثر تطرفًا لهذا.) أما الطريقة الثانية فهي طريقة غير مباشرةٍ؛ فحتى وإن كنت تُؤمن أنَّ «الآلات» ليست هي من تتَّخذ القرارات، بل «الأشخاص الذين صمَّمُوا تلك الآلات وكلَّفوها بمهامها»، فحقيقة أنَّ هؤلاء المُصمِّمين البشريِّين وما فعلوه من تجاهلٍ لأهمية النَّظر إلى الظروف الشَّخصية لكل فردٍ على حدةٍ في تلك الحالات، تُشير إلى أنَّهم قد أعطوا قيمةً ضئيلةً لحياة الآخرين. وقد يكون هذا علامةً على بداية انشقاقٍ عظيمٍ بين النُّخبة الذين يُخدمون بيد البشر، وبين بقيَّة الطبقات المُتدنية الذين تخدمهم الآلات وتتحكَّم فيهم.

في الاتحاد الأوروبي، تحظُر المادة رقم ٢٢ في النظام العام لحماية البيانات لعام ٢٠١٨ بوضوحٍ إعطاء السُّلطة للآلات في الحالات التالية:

> لصاحب البيانات الحق في ألا يخضع لقرارٍ مبني فقط على المُعالجة الآلية؛ بما في ذلك التنميط، الذي يتَّرتَّب عليه آثار قانونية تتعلق به أو تؤثر عليه على نحو ملحوظ.

رغم أنَّ هذا يبدو رائعًا في منطقه، فإنَّنا لا نعرف بعد (على الأقل في وقت كتابة هذا الفصل) مقدار الأثر الذي سيتركه عمليًّا. فغالبًا ما يكون من الأسهل والأسرع والأرخص أن ندع الآلات تتَّخذ القرارات.

وأحد الأسباب التي تدعونا للقلق من القرارات المُؤتمتة هو احتمالية ما يُطلق عليه «انحياز الخوارزميات» — وهو ميل خوارزميات تعلُّم الآلة إلى اتِّخاذ قراراتٍ مُنحازةٍ على نحوٍ غير سليم في أمورٍ مثل القروض والتَّسكين والوظائف والتأمين وإطلاق السَّراح المشروط والعقوبات والتسجيل الجامعي وهلُمَّ جرًّا. والاستناد الصَّريح إلى معايير مثل العرق في هذه القرارات مُجرَّم منذ عقود في العديد من الدول، ومُحظور بنصِّ المادة رقم ٩ من النظام العام لحماية البيانات الخاص بالاتحاد الأوروبي في عددٍ كبيرٍ من التطبيقات. وهذا لا يعني بالطَّبع أنَّ باستبعاد العرق من البيانات، سنحصل بالضرورة على قراراتٍ غير مُنحازةٍ عرقيًّا. على سبيل المثال، بداية من ثلاثينيات القرن الماضي، أقرَّت الحكومة الأمريكية تطبيق مُمارسة التمييز ضدَّ بعض المناطق، والتي تسبَّبت في حرمان بعض الأرقام البريدية من إقراض الرهن العقاري وغيره من أنواع الاستثمار المُختلفة، مما أدي إلى انخفاضٍ في قيمة العقارات. ثمَّ اكتشفنا فجأة أنَّ تلك الأرقام البريدية كان أغلبها لأمريكيين من أصولٍ أفريقية.

ولمنع هذه الممارسة، يُستخدَم الآن أول ثلاثة أرقامٍ من الخمسة الأرقام المكوِّنة للرقم البريدي، لاتِّخاذ القرارات الائتمانية. بالإضافة إلى ذلك، يجب أن تكون عملية اتِّخاذ القرار قابلةً للمُراجَعة للتأكد من عدم وجود أي انحيازاتٍ أخرى «غير مقصودةٍ». يقال عادة إن النظام العام لحماية البيانات الخاص بالاتحاد الأوروبي يُعطي «الحق في التَّفسير» لأي قرارٍ مؤتمت،[38] لكنَّ صياغة المادة رقم ١٤ تتطلَّب فقط ما يلي:

> معلومات مفيدة عن المنطق وراء القرار، وكذلك الأهمية والعواقب المُتوخَّاة من مثل هذه المعالجة لصاحب البيانات.

في الوقت الحاضر، نحن لا نعرف كيف ستطبق المحاكم هذه العبارة وتُدخلها حيِّز التنفيذ. من المُحتمل أن المُستهلك البائس سيُعطى فقط وصفًا لخوارزمية التَّعلُّم المُتعمِّق المُستخدمة في تدريب المُصنِّف الآلي الذي اتَّخذ القرار.

في عصرنا الحالي، تكمُن الأسباب المُحتملة لانحياز الخوارزميات في البيانات نفسها وليس في الانتهاكات المُتعمَّدة من جانب الشركات. في عام ٢٠١٥، أشارت مجلة «جلامور»

إلى اكتشافٍ مُخيِّبٍ للآمال؛ وهو كالتالي: «أول صورةٍ لأنثى عند استخدام خدمة جوجل للبحث في الصور بكلمة CEO تظهر في الصَّف «الثاني عشر» وتُظهر صورةً لدُمية باربي.» (في عام ٢٠١٨، ظهرت بعض صور النساء في نتائج البحث، لكنَّ أغلبهنَّ كُنَّ صورًا عامة جاهزة لسيداتٍ في شكل مديرة تنفيذية، ولكن لم تكن هناك صور حقيقية. في عام ٢٠١٩، كانت النتائج أفضل نوعًا ما.) لم يكن هذا نتيجةً لانحيازٍ مُتعمَّد إلى جنسٍ بعينه في خوارزميات ترتيب الصور في خدمة جوجل للبحث في الصور، لكنَّه كان انحيازًا مُسبقًا في الثقافة التي كانت مصدرًا للبيانات؛ فهناك عدد أكبر بكثيرٍ من المديرين التَّنفيذيِّين من الذكور مقارنةً بالإناث، وعندما يُريد الناس أن يصفوا نموذجًا للمدير التَّنفيذي في صورةٍ ما، فإنهم يختارون دائمًا صورةً لأحد الذكور. وحقيقة أنَّ الانحياز موجود في البيانات في المقام الأول لا يعني بالتأكيد أنَّه لا يُوجَد إلزام لاتِّخاذ بعض الإجراءات لتصحيح المُشكلة.

هناك العديد من الأسباب الأخرى التي يغلب عليها الطابع التِّقني التي قد تدفع بالتَّطبيق البسيط إلى طرق تعلُّم الآلة بأن يُخرج نتائج مُنحازة. على سبيل المثال، الأقلية تُعرَّفُ على أنَّها طائفة لها تمثيل قليل في عينات بيانات سُكان دولةٍ ما؛ ومن ثمَّ، فإن توقُّعات أن يكون الأفراد من الأقليات قد تكون أقلَّ دقَّةً إذا كانت تلك التوقُّعات مُستندةً على نحوٍ كبير على بياناتٍ من أفراد آخرين من نفس المجموعة. ولكن لحُسن الحظ، بُذل قدر كبير من الجهد لحلِّ مُشكلة إزالة الانحياز غير المُتعمَّد من جانب خوارزميات تعلُّم الآلة، وهناك الآن طرُق جديدة لإخراج نتائج غير مُنحازة طبقًا للعديد من التَّعريفات المعقولة والمُستحسنة لمفهوم الإنصاف.[39] والتَّحليل الرياضي لتلك التعريفات لمفهوم الإنصاف يُظهر أنَّها لا يُمكن تحقيقها جميعًا في آنٍ واحد، وأنَّ عند فرض تحقيقها في آنٍ واحد، تتسبَّب في خفض دقة التَّوقعات، وفي حالة اتخاذ قراراتٍ بشأن الإقراض، في ربحٍ أقلَّ للمُقرض. وهذا أمر ربما يكون محبطًا، لكن على الأقل يُوضِّح لنا التَّنازلات اللازمة لتفادي انحياز البيانات. وآمُل أن ينتشر الوعي بهذه الطرق وهذه المُشكلة سريعًا بين صانعي السياسات والمُمارسين والمُستخدِمين.

إذا كان إعطاء الآلات سلطة على أفرادٍ من الجنس البشري قد يُخلِّف بعض المشاكل أحيانًا، فما بالك بإعطائها السُّلطة على جماعاتٍ من البشر؟ بعبارة أخرى، أيجب علينا أن نُعطي للآلات أدوارًا سياسية وإدارية؟ في الوقت الحالي قد يكون هذا التَّصور بعيدًا جدًّا؛ فالآلات لا تستطيع أن تنخرط في محادثاتٍ طويلة وتفتقِر إلى فهم أبسط العوامل

المتعلقة باتِّخاذ القرارات على نطاقٍ واسعٍ؛ مثل: هل ترفع الحد الأدنى للأجور أم لا؟ أو هل ترفض عرض استحواذٍ من شركةٍ أخرى؟ لكن الاتجاه العام واضح كالشمس؛ فالآلات تتخذ قراراتٍ على مستوياتٍ أعلى من التَّحكُّم في العديد من المجالات. لنأخذ شركات الطيران كمثال. في البداية، بدأت أجهزة الكمبيوتر في المساعدة في تنظيم جداول الرحلات. لم يمضِ الكثير من الوقت حتى تولَّت عملية توزيع طواقم الطيران، وحجز المقاعد، وإدارة عمليات الصيانة الدورية. لاحقًا، جرى توصيلها بشبكات المعلومات العالمية لتُوفِّر لمديري شركات الطيران تقارير فورية عن الحالة حتى يستطيعوا التعامُل مع أي مشكلةٍ على نحو فعَّال. أما الآن، فهي تتولَّى مهمة إدارة المشكلات، من إعادة توجيه الطائرات، وإعادة جدولة مواعيد الطواقم، وإعادة حجز المقاعد للمُسافرين ومُراجعة جداول الصيانة.

كلُّ هذا يُعدُّ أمرًا جيدًا من وجهة نظرٍ اقتصاديةٍ لشركات الطيران، ولتجربة المُسافرين. لكن السُّؤال هنا هو ما إذا كانت النظم الحاسوبية ما تزال أدواتٍ في يد البشر، أم أنَّ البشر أصبحوا أدواتٍ في يد النظم الحاسوبية يُغذُّونها بالبيانات ويُصلحون الأخطاء عند الضرورة، لكنهم صاروا لا يفهمون كيف يعمل الأمر بالكامل على أيِّ مستوًى من المستويات. والإجابة تُصبح واضحةً عندما تتعطَّل تلك النُّظُم ونعيش في فوضى عالمية حتى تعود تلك النظم إلى العمل مرة أخرى. مثلًا، في ٣ أبريل ٢٠١٨، تسبَّب انهيارٌ مؤقت في النظام في تأخيرٍ كبير أو إلغاءٍ لحوالي ١٥ ألف رحلة طيرانٍ في أوروبا.[40] وعندما تسبَّبت خوارزميات التداول في الانهيار المُفاجئ عام ٢٠١٠ لبورصة نيويورك، ومحت ١ تريليون دولار في دقائق معدودة، كان الحل الوحيد هو غلق التَّداول. ما حدث حينها لا يزال إلى يومنا هذا غير مفهومٍ بالكامل.

قبل أن تُوجَد أي تقنيةٍ على الأرض، عاش البشر كغيرهم من الحيوانات عيشة الكفاف. لقد وقفنا على أرجلنا، إن جاز التعبير. وبدأ فجر التقنية يبزُغ شيئًا فشيئًا اعتمادًا على هرم من الآلات، وبدأنا نترك بصمتنا كأفرادٍ وكجنسٍ بشري. هناك العديد من الطرائق لتصميم العلاقة بين البشر والآلات؛ فإذا ما صمَّمناها ليظلَّ البشر على قدرٍ كافٍ من الفهم والسُّلطة والاستقلالية، فإن الأجزاء التقنية من هذا النظام يمكن أن تزيد من قدرات البشر زيادةً عظيمةً، مما سيجعل كل واحدٍ منا يقف على قمَّة هرمٍ من المهارات والقدرات، كأنَّه نصف إله إن جاز القول. لكن لننظر بعين الاعتبار إلى العاملة في مستودع متجرٍ إلكتروني. سنرى أنَّها أكثر إنتاجيةً من أسلافها؛ لأنَّ لديها جيشًا صغيرًا من الروبوتات الذين يُحضرون لها حاويات التَّخزين لتلتقط المنتجات منها، لكنها في

الوقت نفسه، تُعدُّ جزءًا من نظامٍ أكبر تتحكَّم فيه خوارزميات ذكية تُقرِّر أين يجب أن تقف تلك العاملة وما هي المنتجات التي عليها أن تلتقطها وتُرسلها للشَّحن. إنها في هذه الحالة تُعتبر نصف مدفونةٍ في ذاك الهرم، وليست واقفةً على قمَّته. وما هي إلا مسألة وقتٍ حتى تملأ الرِّمال ما تبقى من مساحةٍ في الهرم ويختفي دورها للأبد.

الفصل الخامس

الذكاء الاصطناعي الفائق الذكاء

(١) مشكلة الغوريلا

لا يحتاج المرء إلى الكثير من الخيال حتى يُدرك أن جعْل أي شيءٍ أكثر ذكاءً منه يُمكن أن يكون فكرة سيئة. نحن نعرف أن تحكُّمنا في بيئتنا وفي الأنواع الأخرى يرجع إلى ذكائنا، لذا، فإن فكرة وجود شيءٍ آخر أكثر ذكاءً منا — سواء كان إنسانًا آليًّا أو كائنًا فضائيًّا — يُثير في النفس على الفور شعورًا بالقلق.

منذ ما يقرُب من عشرة ملايين عام، أنشأ أسلاف الغوريلا الحديثة (مصادفةً، بالتأكيد) السلالة التي أدَّت إلى ظهور البشر. السؤال الآن: ما شعور الغوريلات تجاه ذلك؟ من المؤكد أنها إن كان بإمكانها أن تتحدَّث عن وضع نوعها الحالي في مقابل البشر، فإنَّ الرأي الذي سيُجمِع عليه أفرادها سيكون في واقع الأمر سلبيًّا جدًّا. إن نوعها ليس له بالأساس أي مستقبل غير الذي يمكن أن نسمح به نحن. ونحن لا نريد أن نكون في وضعٍ مُشابه في مقابل آلات فائقة الذكاء. سأُسمِّي هذا «مشكلة الغوريلا»؛ وهي على وجه التحديد القضية المُتمثلة فيما إذا كان البشر يمكنهم الحفاظ على سيادتهم واستقلاليتهم في عالمٍ يتضمَّن آلاتٍ لدَيها ذكاء أكبر على نحوٍ هائل.

إن تشارلز بابيج وآدا كونتيسة لوفليس، اللذَين صمَّما وكتبا برامج المحرك التحليلي في عام ١٨٤٢، كانا مُدركَين لقدراته الكامنة، لكن بدا أنهما لم يكن لديهما أي هواجس بشأنه.[1] لكن في عام ١٨٤٧، هاجم ريتشارد ثورنتون، محرر «بريميتيف إكسباوندر»، وهي مجلة دينية، بضراوة الآلات الحاسبة الميكانيكية قائلًا:[2]

> إنَّ العقل ... يتجاوز حدوده ويتخلى عن ضرورة وجوده بابتكار آلات تقوم بعمليات «التفكير» المنوطة به ... لكن مَن يعرف إن كانت تلك الآلات، عندما

> نصل بها إلى مرحلة أكبر من الإتقان، قد تفكر في خطة لإصلاح كل عيوبنا ثم تنتج آليًّا أفكارًا تتجاوز حدود عقلنا الفاني!

يُعَدُّ هذا على الأرجح أول تكهُّنٍ بشأن الخطر الوجودي الذي قد نتعرَّض له من جانب الآلات الحاسوبية، لكنه بقي طيَّ النسيان.

في المقابل، طوَّرت رواية صمويل باتلر «إريون»، التي نُشرت في عام ١٨٧٢، الفكرة بعُمقٍ أكبر بكثيرٍ وحقَّقت نجاحًا فوريًّا. إن إريون بلد جرى فيه حظر كل الآلات الميكانيكية بعد حربٍ أهلية مريعة بين مناصري ومعارضي الآلات. يعرض أحد أقسام الرواية والذي يُسمَّى «كتاب الآلات» أصول تلك الحرب ويُقدِّم حجج الطرفَين.[3] وهو يُعَد تنبؤًا مخيفًا للجدل الذي ظهر مرةً أخرى في الأعوام الأولى من القرن الحادي والعشرين.

تتمثَّل حُجة مُعارضي الآلات الأساسية في أنَّ الآلات ستتطوَّر حتى تصِل إلى مرحلة تفقد معها البشرية السيطرة عليها:

> ألسنا بهذا نُوجِد بأيدينا خلفاءنا في قيادة هذه الأرض؟ ألسنا نُضيف يوميًّا إلى جمال وبراعة تنظيمها، ونَهَبُها يوميًّا مهارة أكبر ونُوفِّر لها المزيد والمزيد من تلك القوة التي تجعلها ذاتية الفعل وذاتية التنظيم، والتي ستكون أفضل من أي عقل؟ … في غضون عدة عصور، سنجد أنفسنا الجنس الأدنى …
>
> يجب أن نختار بين تحمل المزيد من المعاناة الحالية ومشاهدة أنفسنا وقد حلَّت محلَّنا تدريجيًّا أشياء من صنع أيدينا، حتى نُصبح بالنسبة لها في مرتبةٍ تُشبه مرتبة حيوانات الحقل بالنسبة لنا. … إن حالة الاستعباد تلك ستتملك منَّا خلسةً وفي هدوء ومن خلال وسائل غير ملحوظة.

إن الراوي أيضًا يسرد الحجة الرئيسية المضادة لمؤيدي الآلات، والتي تستشرف فكرة تكافل الإنسان والآلة التي سنستعرضها في الفصل القادم:

> كانت هناك محاولة جدية واحدة للردِّ على هذا. وقد قال صاحبها إن الآلات كانت ستنتظر لها باعتبارها جزءًا من الطبيعة الجسدية للإنسان، بحيث لن تكون سوى أطراف إضافية بالنسبة له.

على الرغم من أن مناهضي الآلات في إريون كسبوا المعركة، فإن باتلر نفسه يبدو في حيرة من أمره. فمن جانب، يشتكي أن «أهل إريون … سريعون في إبداء حسن التمييز في

محراب المنطق، عندما يظهر فيلسوف من بينهم ويُثير حماستهم من خلال ما يُعرف عنه من امتلاكه لمعرفة خاصة» ويقول: «إنهم قاتلوا بعضهم بسبب مسألة الآلات». وعلى الجانب الآخر، إنه يصف مجتمع إريون بأنه مُتناغِم على نحوٍ ملحوظ ومنتج وحتى مثالي. يتقبل أهل إريون تمامًا حماقة إعادة السير في مسار الابتكار الميكانيكي، وينظرون إلى ما تبقى من الآلات والمحتفظ به في المتاحف «بمشاعر أثري إنجليزي تجاه آثارٍ وثنية أو رءوس أسهم مصنوعة من الحجر الصوان».

من الواضح أن رواية باتلر كانت معروفة لدى آلان تورينج، عندما تأمَّل المستقبل الطويل الأمد للذكاء الاصطناعي في محاضرةٍ ألقاها في مانشستر في عام ١٩٥١:[4]

> يبدو من المُرجَّح أنه بمجرَّد أن يبدأ تطوير تفكير الآلات، فلن يمرَّ وقتٌ طويل حتى يتجاوز قدرات تفكيرنا المحدودة. لن يكون هناك خوف من تقادُم الآلات، وستستطيع التواصُل مع بعضها لشحذ مهاراتها. ولذا، في مرحلة ما، يجبُ أن نتوقَّع أن تكون للآلات السيطرة، بالطريقة الذي يذكُرُها صمويل باتلر في عمله «إريون».

وفي العام نفسه، كرَّر تورينج مخاوفه في محاضرة إذاعية أذيعت عبر أنحاء المملكة المتحدة في المحطة الإذاعية «ثيرد بروجرام» التابعة لهيئة الإذاعة البريطانية:

> إن كان بإمكان أيِّ آلةٍ التفكير، فقد تفكر على نحو أكثر ذكاءً مما نفعل، وحينها، أين سنكون؟ حتى إن استطعنا أن نُبقي الآلات في وضعٍ تابع لنا، على سبيل المثال بإيقاف تشغيلها في اللحظات الحاسمة، فيجب علينا، كنوع، أن نشعُر بإهانة كبيرة. ... إن هذا الخطر الجديد ... بالتأكيد شيء يُمكن أن يُشعرنا بالقلق.

إن مناهضي الآلات من أهل إريون عندما «شعروا بعدم ارتياح شديد تجاه المستقبل»، رأوا أن من «واجبهم كبح جماح الشر، بينما كان لا يزال في استطاعتهم فعل ذلك»، ولذلك، دمَّروا كل الآلات. إن رد فعل تورينج تجاه «الخطر الجديد» و«القلق» هو التفكير في «إيقاف الآلات عن العمل» (على الرغم من أنه سيتَّضح لك عما قريب أن هذا ليس في واقع الأمر خيارًا متاحًا). وفي رواية الخيال العلمي الكلاسيكية التي كتبها فرانك هيربرت «كثيب»، والتي تدور أحداثها في المستقبل البعيد، استطاعت البشرية بشقِّ الأنفس

الانتصار في الحرب الباتلرية، وهي حرب شعواء خاضتْها مع «آلات مفكرة». وحينها، أضيفت وصية جديدة للوصايا العشر؛ وهي: «لا تصنع آلة تُشبه العقل البشري». وتلك الوصية تشمل أيَّ آلات حاسُوبية من أي نوع.

تعكس كلُّ ردود الأفعال المُرعبة هذه المخاوف الأولية التي يُثيرها ذكاء الآلات في النفس البشرية. إن احتمال وجود آلات فائقة الذكاء تجعل المرء يشعر بعدم الراحة. كما أنه من المُمكن منطقيًّا أن تسيطر تلك الآلات على العالم وتُخضع أو تقضي على الجنس البشري. إذا كان لهذا التوجُّه في التفكير أن يستمر، ففي واقع الأمر إن رد الفعل المعقول الوحيد المتاح لنا، في الوقت الراهن، هو محاولة إيقاف الأبحاث في مجال الذكاء الاصطناعي؛ على وجه التحديد، حظر تطوير واستخدام نظم ذكاء اصطناعي عام ويُضاهي الذكاء البشري.

إنني، مثل أغلب الباحثين الآخَرين في مجال الذكاء الاصطناعي، انتفض ذُعرًا من احتمال حدوث هذا. كيف يجرؤ أي شخص على إخباري بما يُمكنني التفكير أو عدم التفكير فيه؟ إن أي شخصٍ يقترح وضع نهاية لمجال الذكاء الاصطناعي يجب أن يقدم الكثير من الحجج المُقنعة المؤيدة لما يريد فعله. إن إغلاق هذا المجال سيعني تجاهُل ليس فقط أحد السبل الرئيسية لفهم طريقة عمل الذكاء البشري، وإنما أيضًا فرصة ذهبية لتحسين وضع البشر؛ وذلك بتطوير حضارة أفضل بكثير. إن القيمة الاقتصادية للذكاء الاصطناعي الذي يُضاهي الذكاء البشري تُقاس بآلاف التريليونات من الدولارات، لذا، فإن الزخم الموجود وراء أبحاث الذكاء الاصطناعي من جانب الشركات والحكومات من المُنتظر أن يكون هائلًا. إنه سيتغلَّب على الاعتراضات الغامضة لأي فيلسوف، مهما بلغ عظم «ما يُعرَف عنه من امتلاكه لمعرفة خاصة»، بحسب تعبير باتلر.

هناك مشكلة أخرى في فكرة حظر أبحاث الذكاء الاصطناعي العام والتي تتمثَّل في صعوبة فعل هذا. يحدث التقدُّم في هذا المجال بالأساس على سبُّورات المعامل البحثية حول العالم، مع ظهور المشكلات الرياضية وحلِّها. نحن لا نعرف مُقدَّمًا أي الأفكار والمعادلات التي يجب حظرها، وحتى لو فعلنا، لا يبدو من المعقول توقُّع أن يكون مثل هذا الحظر مُلزمًا أو فعالًا.

لزيادة الصعوبة أكثر، عادة ما يعمل الباحثون الذين يُحدِثون تقدُّمًا في مجال الذكاء الاصطناعي العام على شيء آخر. كما حاججتُ من قبل، إن البحث في مجال الذكاء الاصطناعي الخاص — تلك الاستخدامات النوعية النافعة مثل لعب الألعاب أو التشخيص

الطبي أو التخطيط للرحلات — عادة ما يُؤدِّي إلى إحراز تقدُّم في تقنيات عامة تكون قابلة للتطبيق في نطاقٍ واسع من الأمور الأخرى ويُقربنا أكثر من الذكاء الاصطناعي الذي يضاهي الذكاء البشري.

لهذه الأسباب، من غير المُحتمَل لِمجتمع الذكاء الاصطناعي — أو الحكومات والشركات التي تتحكم في القوانين والميزانيات البحثية — أن يتعامل مع مشكلة الغوريلا بوقف العمل في مجال الذكاء الاصطناعي. إن كان بالإمكان التعامل مع هذه القضية بهذه الطريقة فقط، فإنها لن تُحل.

إن الطريقة الوحيدة التي يبدو أنها يمكن أن تنجح في حل هذه المشكلة هي فهم سبب إمكانية أن يكون ابتكار ذكاء اصطناعي جيدٍ شيئًا سيئًا. يبدو أننا عرفنا الحل منذ آلاف الأعوام.

(٢) مشكلة الملك ميداس

إن لنوربرت فينر، الذي تحدَّثنا عنه في الفصل الأول، تأثيرًا عميقًا على العديد من المجالات، بما في ذلك الذكاء الاصطناعي والعلوم المعرفية ونظرية التحكُّم. كان فينر، بخلاف معظم مُعاصِريه، مُهتمًّا بوجهٍ خاص بمسألة عدم إمكانية التنبؤ بسلوك النظم المعقدة العاملة في العالم الواقعي. (لقد كتب ورقته البحثية الأولى حول هذا الموضوع وهو في سنِّ العاشرة.) لقد أصبح مُقتنعًا بأن ثقة العلماء والمهندسِين الزائدة في قدرتهم على التحكُّم في ابتكاراتهم، سواء العسكرية أو المدنية منها، يمكن أن تكون لها توابع كارثية.

في عام ١٩٥٠، نشر فينر كتاب «الاستخدام البشري للبشر»،[5] والذي تقول النبذة المكتوبة عنه في غلافه الأمامي «"العقل الميكانيكي" والآلات المُماثلة يمكن أن تُدمِّر القِيَم الإنسانية أو يمكن أن تتيح لنا إدراكها على نحوٍ لم يحدث من قبل».[6] لقد نقح آراءه تدريجيًّا بمرور الوقت وبحلول عام ١٩٦٠، توصَّل لنقطة مهمة وأساسية؛ وهي استحالة تحديد الأهداف البشرية الحقيقية على نحوٍ صحيح وكامل. هذا بدوره يعني أن ما أطلقت عليه النموذج القياسي، الذي يُحاول البشر من خلاله غرس أهدافهم في الآلات، مُقدَّر له الفشل.

يمكن أن نُطلق على هذا «مشكلة الملك ميداس»، وميداس هذا هو ملك أسطوري في الميثولوجيا اليونانية القديمة حصل على ما كان يُريده؛ وهو أن يتحوَّل كلُّ شيءٍ يلمسه إلى ذهب. وقد اكتشف متأخرًا جدًّا أن هذا يسري على طعامه وشرابه وأعضاء أُسرته، ولذا،

مات جوعًا وعطشًا وهو في حالة بؤس شديد. نفس الفكرة سارية في عالم الميثولوجيا البشرية. اقتبس فينر قصة جوته الخاصة بصبي الساحر الذي أمر المكنسة بجلب الماء، لكنه لم يحدد لها كم الماء الذي يريده ولم يكن يعرف كيف يوقف المكنسة.

يمكن صياغة ذلك بطريقة تقنية بأن نقول إننا نعاني من فشل في «توفيق القيم»؛ أي إننا، ربما عن غير قصد، ندمج في الآلات أهدافًا لا تتوافق على نحوٍ تام مع أهدافنا. لقد كنا حتى وقتٍ قريب مَحميين من التوابع الكارثية المُحتملة لذلك من خلال الإمكانيات المحدودة للآلات الذكية والنطاق المحدود لتأثيرها على العالم. (في واقع الأمر، معظم أبحاث الذكاء الاصطناعي تعتمد على مشكلات الألعاب غير الواقعية في المعامل البحثية.) يُعبر فينر عن هذا في كتابه «الرب وجولم» الذي صدر في عام ١٩٦٤ قائلًا:[7]

> في الماضي، لم تكن النظرة الجزئية والمنقوصة للمسعى البشري مُستفزَّة نسبيًّا لأنه صاحَبَها قصور تقني. ... يعدُّ هذا إحدى الحالات العديدة التي حمانا فيها قصورنا البشري من التأثير المُدمر على نحوٍ كامل للحماقة البشرية.

للأسف، انتهت فترة الحماية هذه على نحوٍ سريع.

لقد رأينا بالفعل كيف أن خوارزميات انتقاء المحتوى في مواقع التواصُل الاجتماعي قد أحدثت فوضى في المجتمع بدعوى تعظيم عوائد الإعلانات. وفي حالة كنتَ تعتقد أن تعظيم عوائد الإعلانات كان بالفعل هدفًا حقيرًا ما كان يجب السعي من أجل تحقيقه، فدعنا نفترض بدلًا من ذلك أننا طلبنا من نظامٍ مُستقبليٍّ خارقٍ أن يتبنَّى الهدف النبيل المُتمثل في إيجاد علاجٍ لمرض السرطان؛ على نحوٍ مثالي بأسرع ما يُمكن؛ لأن هناك شخصًا يموت منه كل ٣٫٥ ثانية. في خلال ساعات، سيقرأ نظام الذكاء الاصطناعي الأدبيات الطبية الحيوية بأكملها ويفترض ملايين المُركَّبات الكيميائية التي من المُحتمَل أن تكون فعالة لكنها لم تخضع للاختبار من قبل. وفي خلال أسابيع، سيتسبَّب النظام في إصابة كل البشر بأورامٍ عديدة من أنواع مختلفة حتى يُمكنه عمل تجارب طبية على هذه المركبات؛ نظرًا لأن هذه هي أسرع طريقة لإيجاد علاج لأي مرض. يا للأسف!

إذا كنت تفضل حل مشكلات بيئية، فقد تطلب من الآلة حلَّ مشكلة الزيادة السريعة في نسبة حموضة المحيطات التي تنتج من ارتفاع معدلات ثاني أكسيد الكربون في الغلاف الجوي. ستُطوِّر الآلة مادة مُحفِّزة جديدة تُسهِّل وجود تفاعل كيميائي شديد السرعة بين المحيطات والغلاف الجوي وتُعيد مستويات حموضة المحيطات إلى طبيعتها. للأسف،

سيُستهلك ربع الأكسجين الموجود في الغلاف الجوي في هذه العملية، مما سيجعلنا نتنفَّس بصعوبةٍ وعلى نحوٍ مؤلم. يا للأسف!

إن تلك الأنواع من سيناريوهات نهاية العالم واضحة؛ كما قد يتوقع المرء فيما يتعلق بتلك السيناريوهات. لكن هناك العديد من السيناريوهات التي فيها نوع من الاختناق العقلي «يتسرَّب إلينا خلسة في هدوء وبطرُق غير ملحوظة». إن مقدمة كتاب ماكس تيجمارك «الحياة ٣٫٠» تصِف ببعض التفصيل سيناريو تتحكَّم فيه آلة فائقة الذكاء تدريجيًّا من الناحية الاقتصادية والسياسية في العالم بأكمله دون أن يكتشف ذلك أحد. إن الإنترنت والآلات ذات النطاق العالمي التي تدعمها — تلك التي تتفاعل بالفعل مع مليارات «المستخدمين» على نحو يومي — توفر البيئة المثالية لتحكُّم الآلات في البشر.

أنا لا أتوقَّع أن يكون الهدف الذي سيُدمَج في تلك الآلات من النوعية التي تتضمن مسألة «السيطرة على العالم». من المُحتمَل أكثر أن يكون تعظيم الأرباح أو تعظيم المشاركة أو ربما حتى هدف يبدو محمودًا مثل تحقيق درجات أعلى في الاستبيانات المُنتظمة الخاصة بمدى سعادة المُستخدِمين، أو تقليل استخدامنا للطاقة. والآن، إذا كنا نرى أنفسنا كيانات فِعالنا يُتوقَّع منها أن تُحقِّق غاياتنا؛ فهناك طريقتان لتغيير سلوكنا. الأولى هي الطريقة العتيقة الطراز؛ والمُتمثلة في ترك توقعاتنا وأهدافنا دون تغيير، وتغيير ظروفنا المحيطة؛ على سبيل المثال، بأن يُعرض علينا المال أو نتعرَّض للتهديد أو التجويع حتى نتقبَّل التغيير. وهذا أمر مُكلِّف وصعب بحيث يصعُب على أي جهاز كمبيوتر فعله. أما الطريقة الثانية، فتتمثَّل في تغيير توقُّعاتنا وأهدافنا. هذا أسهل بكثيرٍ بالنسبة إلى أي آلة. فهي على تواصُل معك لعدة ساعات كلَّ يومٍ وتتحكَّم في وصولك للمعلومات وتُوفِّر لك معظم احتياجاتك من الترفيه من خلال الألعاب وبرامج التلفزيون والأفلام والتفاعُل الاجتماعي.

ليس لدى خوارزميات التعلُّم المُعزَّز التي تُحسِّن معدل النقر في وسائل التواصل الاجتماعي القدرة على التفكير على نحوٍ منطقي في السلوك البشري؛ في واقع الأمر، إنها حتى لا تعرف بأي نحوٍ بوجود البشر من الأساس. بالنسبة للآلات التي لدَيها فَهم أكبر للجوانب النفسية والمُعتقدات والدوافع البشرية؛ فيجِب أن يكون من السهل نسبيًّا أن تُرشدنا تدريجيًّا في اتجاهاتٍ تزيد من درجة تحقيقها لأهدافها. على سبيل المثال، قد تُقلل من استهلاكنا للطاقة بإقناعنا بأن يكون لدَينا عددٌ أقل من الأبناء، مما يُحقق في النهاية

— وعن غير قصد — أحلام الفلاسفة المؤيدين لتحديد النسل، الذين يرغبون في تقليل الأثر المُدمر للبشرية على العالم الطبيعي.

مع بعض المُمارسة، يُمكنك تعلم كيفية تحديد الطرق التي يُمكن من خلالها تحقيق أي هدفٍ مُحدد بنحوٍ أو بآخر أن يُؤدِّي إلى عواقب وخيمة. ويتضمَّن أحد الأنماط الأكثر شيوعًا في هذا الشأن حذف شيءٍ من الهدف لا تهتمُّ به بالفعل. في تلك الحالات — كما في الأمثلة السالفة الذكر — سيجد في الغالب نظام الذكاء الاصطناعي حلًّا مثاليًّا يتعامل مع الشيء الذي تهتمُّ به بالفعل، ولكني نسيتُ أن أقول إنه سيفعل ذلك على نحوٍ مُبالَغ فيه. لذا، إذا قُلت لسيارتك الذاتية القيادة: «خُذيني إلى المطار بأسرع ما يمكن!» وفسَّرت هي ذلك على نحوٍ حرفي، فستصل إلى سرعةٍ قدرها ١٨٠ ميلًا في الساعة وستدخُل أنت السجن. (لحُسن الحظ، لن تقبل السيارات الذاتية القيادة الموجودة حاليًّا مثل هذا الطلب.) إذا قلت: «خُذيني إلى المطار بأسرع ما يُمكن دون تجاوز حدِّ السرعة المتعارف عليه»، فإنها ستُسرع وتتوقف بأقصى ما يُمكنها، وتناور داخل حالات الاختناق المروري وخارجها للحفاظ على الحد الأقصى للسرعة فيما بينهما. وقد تُزيح حتى السيارات الأخرى من طريقها لتربح بضع ثوانٍ في الحشد الفوضوي الموجود أمام مبنى الركاب بالمطار. وهكذا سيستمر الأمر بحيث، في النهاية، تكون عليك إضافة اعتبارات أخرى كافية بحيث تقترِب قيادة السيارة على نحوٍ كبير من تلك الخاصَّة بسائقٍ بشري ماهر يأخذ شخصًا إلى المطار على نحو سريع.

إن القيادة مُهمة بسيطة ذات تبعات محلية فقط، كما أن نظم الذكاء الاصطناعي المُستخدمة حاليًّا في مجال القيادة ليست ذكية جدًّا. لهذَين السببَين، يمكن توقع العديد من أنماط الفشل المحتملة؛ وستكشف أنماط أخرى عن نفسها من خلال نُظُم المحاكاة الخاصة بالقيادة أو ملايين الأميال من الاختبار مع سائقين محترفين مُستعدِّين لتولي القيادة في حالة حدوث خطأ؛ في حين ستظهر أُخرى لكن لاحقًا عندما تكون السيارات بالفعل على الطريق ويحدث شيء غريب.

لسوء الحظ، في حالة النظم الخارقة الذكاء التي يُمكن أن يكون لها تأثير عالمي، لا تُوجَد نظم محاكاة ولا فُرَص لتصحيح الأوضاع. وبالتأكيد، من الصعب للغاية، وربما من المُستحيل، للبشر أن يتوقَّعوا ويستبعدوا مُقدَّمًا كل الطرق المدمِّرة التي يُمكن أن تختارها الآلة لتحقيق هدفٍ مُعيَّن. بوجه عام، إذا كان لديك هدف ولآلة خارقة الذكاء هدف آخر مختلف ومُتعارِض، فإنَّ الآلة ستحصُل على ما تُريد، أما أنت، فلا.

(٣) الخوف والحقد: الأهداف الأداتية

إن بدا أنَّ وجود آلة تتبع هدفًا غير صحيح شيءٌ سيئ بالقدر الكافي، فإنَّ هناك ما هو أسوأ من ذلك. إن الحل الذي اقترحه آلان تورينج — وهو إيقاف التشغيل في اللحظات الحاسمة — قد لا يكون مُتاحًا، لسبب بسيط جدًّا؛ وهو أنك «لا يُمكنك جلب فنجان القهوة إذا كنت ميتًا».

دعْني أوضح لك الأمر. افترض أن آلةً هدفها هو جلب القهوة. إذا كانت ذكية بالقدر الكافي، فإنها ستفهم بالتأكيد أنها ستفشل في تحقيق هدفها إذا توقَّفَت عن العمل قبل إكمال مهمَّتها. ومن ثمَّ فإن هدف جلب القهوة ينشئ هدف تعطيل زرِّ الإغلاق، باعتباره هدفًا فرعيًّا ضروريًّا. وينطبق الأمر نفسه على هدف علاج السرطان أو حساب أرقام ثابت الدائرة. لا يُوجَد في واقع الأمر الكثير الذي يُمكنك فعله بمجرد أن تموت، لذا، يمكنك أن تتوقَّع أن تتصرَّف نظم الذكاء الاصطناعي على نحو استباقي للحفاظ على وجودها، مع الوضع في الاعتبار امتلاكها لأي هدفٍ مُحدَّد بنحوٍ أو بآخر.

إذا تعارض هذا الهدف مع التفضيلات البشرية؛ فلدَينا بالضبط ما حدث في حبكة فيلم «٢٠٠١: ملحمة الفضاء» («٢٠٠١: آه سبيس أوديسي») التي قتل فيها الكمبيوتر هال ٩٠٠٠ أربعةً من رواد الفضاء الخمسة الذين كانوا على متن سفينة فضاء لمنع تداخُلهم مع مُهمَّته. استطاع رائد الفضاء الأخير المُتبقِّي، ديف، إيقاف تشغيل هذا الكمبيوتر بعد معركةٍ عقلية ملحمية؛ على الأرجح كي يُحافظ أصحاب الفيلم على جاذبية الحبكة. لكن إذا كان هال خارق الذكاء حقًّا، ما كان سيستطيع ديف إيقاف تشغيله على الإطلاق.

من المهم معرفة أن الحفاظ على الذات لا يجِب أن يكون نوعًا من الغريزة الداخلية أو الدافع الأساسي في الآلات. (لذا، القانون الثالث لعلم الروبوتات[8] الذي وضعه إيزاك أزيموف الذي يبدأ بالآتي: «يجب على الروبوت أن يحمي وجوده» غير ضروري بالمرة.) فلا حاجة إلى دمج هدف الحفاظ على الذات في أيِّ آلة لأنه «هدف أداتي»؛ وهو هدف فرعي مفيد تقريبًا لأي هدف رئيسي.[9] إن أي كيانٍ لدَيه هدف مُحدَّد سيتصرَّف تلقائيًّا كما لو أن له أيضًا أهدافًا أداتية.

إنَّ امتلاك المال يُعدُّ هدفًا أداتيًّا داخل نظامنا الحالي، بالإضافة إلى الاستمرار في العمل. لذا، قد تحتاج أي آلة ذكية إلى المال، لا لأنها جشعة ولكن لأن المال مُفيد في تحقيق كافة أنواع الأهداف. في فيلم «التسامي»، عندما حُمل عقل جوني ديب في الكمبيوتر الفائق الكمي، فإن أول شيءٍ فعلته الآلة هو نسخ نفسها على ملايين أجهزة الكمبيوتر الأخرى

على الإنترنت حتى لا يمكن لأحدٍ إيقاف تشغيلها. وثاني شيءٍ فعلته هو تحقيق أرباح كبيرة في البورصة لتمويل خطط التوسُّع الخاصة بها.

ماذا كانت خطط التوسُّع تلك على وجه التحديد؟ إنها تتضمَّن تصميم وإنشاء كمبيوتر خارق كمي أكبر بكثيرٍ والقيام بأبحاثٍ في مجال الذكاء الاصطناعي واكتشاف معلومات جديدة في الفيزياء وعلم الأعصاب والبيولوجيا. إن تلك الأهداف الخاصة بالمصادر — القوة الحاسوبية والخوارزميات والمعرفة — هي أيضًا أهداف أداتية مُفيدة في تحقيق أي هدف شامل.[10] إنها تبدو غير ضارة حتى يُدرك المرء أن عملية الاكتساب ستستمرُّ بلا حدود. ويبدو أن هذا سيُوجِد صراعًا حتميًّا مع البشر. وبالطبع، الآلة، المُزوَّدة بنماذج أفضل دائمًا لصنع القرار البشري، ستتوقَّع كل تحرُّك لنا في هذا الصراع وتقضي عليه.

(٤) انفجارات الذكاء

كان آي جيه جود رياضيًّا بارعًا يعمل مع آلان تورينج في حديقة بلتشلي في فك الشفرات العسكرية الألمانية أثناء الحرب العالمية الثانية. وقد تشارك مع آلان اهتماماته الخاصة بذكاء الآلات والاستدلال الإحصائي. وفي عام ١٩٦٥، كتب ما يُعد الآن بحثه الأشهر «تكهنات بشأن أول آلة فائقة الذكاء».[11] تشير أول جملة في البحث إلى أن جود، المُنزعج بسبب الأزمة النووية التي كانت على وشك الانفجار في الحرب الباردة، كان يرى أن الذكاء الاصطناعي يُعدُّ مُنقذًا مُحتملًا للبشرية: «يعتمد بقاء الإنسان على البناء المُبكِّر لآلةٍ فائقة الذكاء». لكنه أثناء عرضه أصبح أكثر تحفُّظًا. وقدم مفهوم «انفجار الذكاء»، لكنه، شأنه شأن باتلر وتورينج وفينر من قبله، كان قلقًا بشأن فقدان السيطرة:

> يُمكن تعريف الآلة الفائقة الذكاء بأنها آلة يُمكن أن تتفوَّق على أيِّ شخصٍ مهما كانت درجة ذكائه في أداء كل الأنشطة العقلية الخاصة به. وحيث إن تصميم الآلات يُعد أحد هذه الأنشطة العقلية، فإن الآلة الفائقة الذكاء يُمكنها حتى تصميم آلات أفضل؛ سيكون هناك حينها بلا شك «انفجار ذكاء»، وسيتخلَّف ذكاء البشر بشدَّة عن الركب؛ ومن ثَمَّ ستكون أولى الآلات الفائقة الذكاء هي آخر ابتكارٍ يحتاج الإنسان لوضعه، بشرط أن تكون الآلة طيِّعةً بالقدر الكافي بحيث تُخبرنا كيف نُبقيها تحت السيطرة. من الغريب أن تلك النُّقطة نادرًا ما تُثار خارج نطاق أدب الخيال العلمي.

تُعَدُّ تلك الفقرة عماد أي نقاشٍ حول الذكاء الاصطناعي الخارق، على الرغم من أنَّ التحذير الوارد في نهايتها عادةً ما يجري تجاهُلُه. إن فكرة جود يُمكن تأكيدها بملاحظة أن الآلة الفائقة الذكاء يمكنها ليس فقط تحسين تصميمها، وإنما من المُحتمَل أنها ستفعل ذلك لأنَّ أيَّ آلةٍ ذكية، كما رأينا، تتوقَّع الاستفادة من تحسين مُكوِّناتها المادية وبرامجها. إن احتمالية حدوث انفجار ذكاء عادة ما يجري اقتباسُها باعتبارها المصدر الأساسي للخطر على البشرية من جانب الذكاء الاصطناعي لأنها ستُعطينا وقتًا قليلًا جدًّا لحلِّ مُشكلة التحكم.[12]

إن مُحاجَّة جود بالتأكيد لها وجاهة في ضوء القياس الطبيعي للانفجار الكيميائي الذي فيه يُطلِق كل تفاعل جُزيئي طاقة كافية لبدء المزيد من التفاعلات. على الجانب الآخر، من المُمكن منطقيًّا أن تكون هناك نتائج تناقُصية للتحسينات الخاصة بالذكاء، بحيث تتضاءل تدريجيًّا العملية بدلًا من أن تنفجر.[13] لا تُوجَد طريقة واضحة لإثبات أن عملية الانفجار ستحدث «بالضرورة».

إن سيناريو النتائج التناقُصية مثير للاهتمام في حد ذاته. إنه يُمكن أن ينشأ إذا اتَّضح أن تحقيق نسبة مُعينة من التحسين أصبح أصعب مع ازدياد ذكاء الآلة. (أنا أفترض من أجل المُحاجَّة فقط أن ذكاء الآلة العام قابل للقياس باستخدام نوع مُعيَّن من المقاييس الخطية، وهو ما أشكُّ أنه سيتحقَّق يومًا ما.) في تلك الحالة، لن يتمكَّن البشر أيضًا من بناء ذكاء خارق. إن استنفدت أي آلة خارقة الذكاء بالفعل طاقتها أثناء محاولتها تحسين ذكائها، فإن البشر سيحدث لهم ذلك قبلَها بكثير.

صحيح أنني لم أسمع قطُّ أي حجةٍ قوية مفادها أن بناء أي مستوًى مُعين من ذكاء الآلة ببساطة ليس في استطاعة الذكاء البشري، لكني أفترض أن المرء يجب أن يُقرَّ بأن هذا مُمكن منطقيًّا. «إن هذا مُمكن منطقيًّا» و«أنا على استعداد لرهن مستقبل الجنس البشري على هذا» هما أمران، بالطبع، مختلفان تمامًا. فإن الرهان ضد الذكاء البشري يبدو رهانًا خاسرًا.

إن حدث بالفعل انفجار ذكاء، ولم نستطع حينها حلَّ مشكلة التحكم في الآلات التي لدَيها ذكاء خارق محدود فقط — على سبيل المثال، إذا لم نستطع منعها من إجراء تلك التحسينات الذاتية المُتكررة — فلن يكون لدَينا وقت لحلِّ مُشكلة التحكُّم وسينتهي الأمر. هذا هو سيناريو «التطوُّر السريع» الذي طرحه بوستروم، والذي فيه ذكاء الآلة سيتطور على نحوٍ خياليٍّ في غُضون أيام أو أسابيع. وبعبارة تورينج، إنه «بالتأكيد شيء يُمكن أن يُشعِرنا بالقلق».

يبدو أن ردود الأفعال المُمكنة تجاه هذا القلق ستتمثَّل في عدم الاستمرار في الأبحاث الخاصة بالذكاء الاصطناعي، وإنكار وجود مخاطر خفية في تطوير ذكاء اصطناعي مُتقدِّم، وفهم تلك المخاطر من خلال تصميم نُظم ذكاء اصطناعي تبقى بالضرورة تحت السيطرة البشرية والتقليل من حدتها، والانسحاب؛ ببساطة، ترك المستقبل للآلات الذكية.

إن الإنكار والتقليل من تأثير مخاطر الذكاء الاصطناعي الخارق هما موضوعا ما تبقَّى من هذا الكتاب. وكما حاججتُ من قبل، إن إيقاف البحث في مجال الذكاء الاصطناعي غير مُحتمَل الحدوث (لأنَّ الفوائد المتروكة كبيرة جدًّا) ومن الصعب جدًّا تحقيقه. يبدو الانسحاب أسوأ ردود الأفعال المُمكنة. إنه عادة ما يُصاحِبه فكرة أنَّ نظم الذكاء الاصطناعي الأكثر ذكاءً منَّا على نحوٍ ما «تستحق» أن ترث الكوكب، تاركة للبشر الاستسلام للوضع، ويكون عزاؤهم الوحيد في ذلك هو فكرة أن نسْلَهم الإلكتروني الذكي مُنشغِل بتحقيق أهدافه. لقد نشر تلك الفكرة عالم المستقبليات والمُتخصِّص في علم الروبوتات هانس مورافيك[14] الذي كتب يقول: «سيمتلئ العالم الإلكتروني الهائل بالعقول الفائقة الذكاء غير البشرية المُنشغلة بأمورٍ غير مهمَّة للبشر كما أن أمور البشر غير مهمة للبكتيريا». يبدو أن هذا خطأ. فالقيمة، بالنسبة إلى البشر، تُحدِّدها بالأساس تجربة بشرية واعية. وإذا لم يكن هناك بشر ولا كيانات واعية أخرى تجربتها الذاتية مُهمَّة لنا، فلن تكون هناك أي قيمة.

الفصل السادس

الجدل غير الواسع الدائر حول الذكاء الاصطناعي

«إن تبعات إدخال نوعٍ ذكي آخر إلى الأرض، بعيدة المدى بالقدر الكافي بحيث لا تستحق التفكير الجدي».[1] هكذا أنهت مجلة «ذي إيكونوميست» مراجعتها النقدية لكتاب نيك بوستروم «الذكاء الخارق». إن أغلبنا سيرَون هذا باعتباره مثالًا كلاسيكيًّا على التهوين البريطاني للأمور. أنت بالتأكيد قد تعتقد أن العقول الكبيرة في الوقت الحاضر تقوم بالفعل بهذا التفكير الجدي؛ أي إنها مُنخرطة في نقاشٍ جادٍّ وتوازُن بين المخاطر والفوائد وتبحث عن حلولٍ وتُفتش عن الثغرات الموجودة في الحلول وهكذا. إن الأمر لم يصِل إلى هذا الحد بعدُ، بحسب علمي.

عندما يُقدِّم شخص لأول مرة تلك الأفكار لجمهور مُتخصص في المجال التقني، يستطيع أن يرى فقاعات الأفكار تنبثق من رءوسهم، والتي تبدأ بالكلمات «لكن، لكن، لكن ...» وتنتهي بعلامات تعجُّب.

يأخذ أول نوع من كلمة «لكن» شكل الإنكار. يقول المنكرون: «لكن تلك لا يمكن أن تكون مشكلة حقيقية؛ لأن كذا كذا». بعض هذه الأسباب تعكس تفكيرًا يمكن وصفه بالتفكير التوَّاق، في حين أن البعض الآخر يكون أكثر وجاهة. أما النوع الثاني من كلمة «لكن» فيأخذ شكل التهرُّب؛ أي قبول أن المشكلات حقيقية لكن الزعم بأننا يجب ألا نُحاول حلَّها، إما لأنها غير قابلة للحل وإما لأن هناك أمورًا أخرى أكثر أهمية علينا أن نُركز عليها من نهاية العالم، وإما لأنه من الأفضل ألا نهتمَّ بها على الإطلاق. أما النوع الثالث من كلمة «لكن»، فيأخذ شكل حلٍّ فوري مُبسَّط: «لكن أليس من المُمكن أن نقوم فقط بكذا كذا؟» وكما هو الوضع في حالة الإنكار، بعض هذه الحلول تكون غير

مُجدية على نحوٍ واضح. في حين تقترب أخرى، على الأرجح بالصدفة، من تحديد الطبيعة الحقيقية للمشكلة.

أنا لا أقصد الإشارة إلى أنه لا يُمكن أن تُوجَد أي اعتراضاتٍ مقبولة على فكرة أن الآلات الخارقة السيئة التصميم ستُشكل خطرًا كبيرًا على البشرية. المسألة أنني لم أرَ حتى الآن أيًّا من تلك الاعتراضات. وحيث إن الأمر يبدو على قدرٍ كبير من الأهمية؛ فهو يستحقُّ نقاشًا عامًّا على أعلى مستوى. لذا، من أجل تعزيز ذلك النقاش، وعلى أمل إشراك القارئ فيه، دعوني أُقدم لكم نظرةً سريعة على أبرز ما تمَّ في هذا الشأن حتى الآن، دون تجميل.

(١) الإنكار

إن أسهل طريقة للتعامُل مع الأمر هي إنكار وجود مشكلة من الأساس. بدأ سكوت ألكسندر، صاحب مدونة «سليت ستار كودكس»، مقالًا شهيرًا عن مخاطر الذكاء الاصطناعي كما يلي:[2] «لقد بدأتُ لأول مرة الاهتمامَ بمخاطر الذكاء الاصطناعي تقريبًا في عام ٢٠٠٧. في ذلك الوقت، كان رد فعل معظم الناس تجاه هذا الموضوع هو: «ها ها، عُد عندما يؤمن أي شخص بهذا إلى جانب مُستخدمي الإنترنت الغريبي الأطوار والعشوائيِّين.»»

(١-١) الملاحظات غير المُجدية على نحوٍ واضح

إن أي تهديدٍ مُتصوَّر للمهنة التي يعمل بها أي شخص طوال حياته يُمكن أن تقوده، حتى لو كان ذكيًّا للغاية وعميق التفكير في أغلب الأحيان، إلى أن يقول أشياء قد يرغب في التراجُع عنها والتبرُّؤ منها عند القيام بتحليلٍ أكبر للموضوع ذي الصلة. ونظرًا لأن هذا هو الوضع، فلن أذكُر أصحاب الحجج التالية، الذين جميعهم من الباحثين المعروفين في مجال الذكاء الاصطناعي. لقد ضمنت تفنيدًا لتلك الحجج، حتى لو كان ذلك غير ضروري على الإطلاق.

- الحاسبات الإلكترونية تتفوَّق على البشر في العمليات الحسابية. وحيث إن تلك الآلات لم تُسيطر على العالم، فلا داعيَ للقلق من الذكاء الاصطناعي الخارق.

– التفنيد: الذكاء يختلف عن إجراء العمليات الحسابية، والقدرات الحسابية للحاسبات لا تُتيح لها السيطرة على العالم.

- الخيول لدَيها قوة تفوق البشر، ونظرًا لأننا لا نخشى خروجها عن السيطرة، فلا نحتاج إلى القلق من خروج نُظُم الذكاء الاصطناعي عن السيطرة.

  - التفنيد: الذكاء يختلف عن القوة البدنية، وقوة الخيول لا تُتيح لها السيطرة على العالم.

- تاريخيًّا، لا تُوجَد أي سوابق لقتل الآلات لملايين البشر، لذا، نستدلُّ من ذلك على أن هذا لا يمكن أن يحدث في المستقبل.

  - التفنيد: يُمكن أن يحدث أي شيء، دون أن تكون له سوابق من قبل.

- لا يُمكن لأي كمية مادية في الكون أن تكون لا نهائية، وهذا يتضمَّن الذكاء، لذا، المخاوف من الذكاء الاصطناعي الخارق مُبالَغ فيها.

  - التفنيد: الذكاء الاصطناعي الخارق لا يحتاج لأن يكون لا نهائيًّا حتى يُسبب مشكلات، والفيزياء تسمح ببناء أجهزة حاسوبية أقوى من العقل البشري بمليارات المرات.

- نحن لا نقلق من الأمور التي يقلُّ احتمال حدوثها على نحوٍ كبير، والتي قد تؤدِّي إلى فناء الأنواع؛ مثل ظهور الثقوب السوداء عند المدار الأقرب إلى الأرض، لذا، لِمَ القلق من الذكاء الاصطناعي الخارق؟

  - التفنيد: إذا كان معظم الفيزيائيين على كوكبنا يعملون على صُنع مثل هذه الثقوب السوداء، ألن نسألهم إن كان ذلك لا يُمثِّل أي خطر؟

(١-٢) الأمر معقد

من الأمور الرئيسية في علم النفس الحديث أنَّ أيَّ مُعدَّل للذكاء لا يُمكنه وصف الثراء التام للذكاء البشري.[3] هناك، كما تقول النظرية، أبعاد مختلفة للذكاء؛ سواء المكاني أو المنطقي أو اللغوي أو الاجتماعي أو غير ذلك. ربما كان لأليس، لاعبة كرة القدم التي عرضنا لها في الفصل الثاني، ذكاء مكاني أكبر من صديقها بوب، ولكنَّ ذكاءها الاجتماعي أقل منه. لذا، لا يُمكننا ترتيب كل البشر على نحوٍ محكم فيما يتعلق بالذكاء.

هذا حتى ينطبق أكثر على الآلات لأن قدراتها أقل منَّا بكثير. إن محرك البحث الخاص بجوجل وبرنامج «ألف جو» ليس بينهما تقريبًا أي شيءٍ مُشترك، هذا بالإضافة إلى كونهما مُنتجين لشركتَين فرعيَّتَين تنتميان لنفس الشركة الأم، لذا، لا داعي للقول بأنَّ أحدهما أكثر ذكاءً من الآخر. وهذا يجعل مفاهيم «معدَّل ذكاء الآلات» مُلغِزة، ويُشير إلى أنه من المُضلِّل وصف المستقبل باعتباره سباقًا أحاديَّ البُعد فيما يتعلَّق بمعدل الذكاء بين البشر والآلات.

طوَّر كيفين كيلي، المُحرِّر المؤسس لمجلة «وايرد» والمعلق المُتبصِّر على نحوٍ ملحوظ في المجال التقني، هذه الفكرة أكثر. ففي مقاله «خرافة الذكاء الاصطناعي الخارق»،[4] كتب يقول: «الذكاء ليس له بُعد واحد، لذا، فإن مفهوم «أذكى من البشر» لا معنى له». وهكذا، وبضربةٍ واحدة، جرى تبديد كل المخاوف بشأن الذكاء الخارق.

أحد الردود الواضحة على ذلك هو أن الآلة يُمكنُها تجاوُز القدرات البشرية في «كل» أبعاد الذكاء ذات الصلة. وفي هذه الحالة، حتى بمعايير كيلي الصارمة، ستكون الآلة أكثر ذكاءً من الإنسان. لكن هذا الافتراض القوي ليس ضروريًّا لتفنيد حجة كيلي. تأمَّل معي حيوانات الشمبانزي. ربما يكون لدى هذه الحيوانات ذاكرة مدى قصير أفضل من البشر، حتى في المهام البشرية الطابع مثل تذكُّر تسلسلات من الأرقام.[5] إن ذاكرة المدى القصير بُعد مُهمٌّ للذكاء. ومن ثمَّ، وبالنظر إلى حُجة كيلي، البشر ليسوا أذكى من حيوانات الشمبانزي؛ في واقع الأمر، سيزعُم هو أن مفهوم «أذكى من الشمبانزي» لا معنى له. إن هذا يُعطي بعض العزاء لحيوانات الشمبانزي (وحيوانات البونوبو والغوريلا وإنسان الغاب والحيتان والدلافين وما إلى ذلك) حيث إنها أنواع تعيش فقط لأننا تكرَّمنا بالسماح لها بذلك. وهو لا يُعطي أي عزاءٍ لكل تلك الأنواع التي تسبَّبنا بالفعل في محوِها من على وجه الأرض. وهو أيضًا يُعطي بعض العزاء للبشر الذين قد يقلقون من أن تحلَّ محلهم الآلات.

(١-٣) الأمر مستحيل

حتى قبل ظهور مجال الذكاء الاصطناعي في عام ١٩٥٦، كان المُفكرون العظام يتشكَّكون ويقولون إن صُنع آلاتٍ ذكية أمر مستحيل. خصَّص آلان تورينج معظم بحثه الشهير الذي ظهر في عام ١٩٥٠ والذي كان بعنوان «الآلات الحاسوبية والذكاء» لتنفيذ تلك الحجج. ومنذ ذلك الحين، أخذ مجتمع الذكاء الاصطناعي يُفنِّد مزاعم مُماثلة من جانب الفلاسفة،[6] وعلماء الرياضيات،[7] وغيرهم. وفي الجدل الدائر حاليًّا حول الذكاء الخارق، أطلق العديد

من الفلاسفة مزاعم الاستحالة هذه ليثبوا أن البشرية ليس لديها ما تخشاه.[9،8] وهذا ليس أمر مُفاجئًا.

إن «دراسة المائة عام حول الذكاء الاصطناعي» هي مشروع طموح طويل الأجل ترعاه جامعة ستانفورد. وهدف تلك الدراسة هو تتبُّع الذكاء الاصطناعي، أو، بالأحرى، «دراسة وتوقع كيف ستتسلَّل آثار الذكاء الاصطناعي عبر كل جوانب عمل البشر وحياتهم ولعبهم». وكان أول تقرير رئيسي لتلك الدراسة والذي جاء بعنوان «الذكاء الاصطناعي والحياة في عام ٢٠٣٠» مفاجأة.[10] فكما قد يكون مُتوقَّعًا، إنه يؤكد على فوائد الذكاء الاصطناعي في مجالات مثل التشخيص الطبي وأمان المَركبات. لكن الشيء غير المتوقع هو الزعم بأنه «لا تُوجَد سلالة من الروبوتات الخارقة في الأفق ولا حتى هذا مُمكن، وذلك بخلاف ما يظهر في الأفلام السينمائية».

بحسب معلوماتي، هذه هي المرة الأولى التي يتبنَّى فيها على نحوٍ علني باحثون جِدِّيون في مجال الذكاء الاصطناعي وجهة النظر القائلة بأن الذكاء الاصطناعي الخارق أو ذلك الذي يُضاهي الذكاء الإنساني مُستحيل، وهذا يحدث وسط فترةٍ يحدث فيها تطور شديد السرعة في الأبحاث في هذا المجال، مع انهيار الحواجز الواحد تلو الآخر. يبدو الأمر كما لو أن مجموعة من كبار علماء البيولوجيا العاملين في مجال أمراض السرطان أعلنوا أنهم كانوا يخدعوننا طوال الوقت؛ فلطالما كانوا على يقينٍ بأنه لن يكون هناك أبدًا علاج للسرطان.

ترى، ماذا قد يكون وراء هذا التغيُّر الكامل والمُفاجئ؟ لا يُقدِّم التقرير أي حجج أو أدلة على الإطلاق. (في واقع الأمر، ما الأدلة التي يُمكن تقديمها على استحالة ظهور ترتيبٍ مُعين من الذرات يُمكنه التفوُّق على العقل البشري؟) أشك أنَّ هناك سببَين. الأول: هو الرغبة الطبيعية لنفي وجود مشكلة الغوريلا، والتي تُمثِّل احتمالًا غير مريح بالمرة للباحث في مجال الذكاء الاصطناعي؛ بالتأكيد، إذا كان الذكاء الاصطناعي الذي يُضاهي الذكاء البشري مستحيل الوجود، فستختفي مشكلة الغوريلا على نحوٍ رائع. أما الثاني، فيتمثل في «القبلية»؛ أي الميل إلى اتخاذ موقف دفاعي ضد ما يُرى على أنه «محاولات للنيل» من الذكاء الاصطناعي.

يبدو من الغريب النظر إلى الزعم بأن وجود الذكاء الاصطناعي الخارق مُمكن باعتباره محاولة للنيل من مجال الذكاء الاصطناعي، ويبدو حتى أغرب الدفاع عن الذكاء الاصطناعي بالقول بأنه لن ينجح أبدًا في تحقيق أهدافه. نحن لا يمكننا حماية أنفسنا من احتمال حدوث كوارث مستقبلية فقط بالرهان على عدم براعة العبقرية البشرية.

لقد قُمنا بتلك الرهانات من قبل وخسرنا. فكما أوضحنا من قبل، إن كبار علماء الفيزياء في أوائل ثلاثينيات القرن الماضي، والذين يُمثِّلهم اللورد رذرفورد، كانوا يعتقدون بثقة أن إنتاج الطاقة النووية مُستحيل، لكن ابتكار ليو سلارد التَّفاعلَ النَّووي المُتسلسل المُستحَثَّ بالنيوترونات في عام ١٩٣٣ أثبت أن تلك الثقة ليست في محلها.

إن الإنجاز الذي حقَّقه سلارد جاء في توقيتٍ صعب؛ إذ جاء مع بداية سباق تسلُّح مع ألمانيا النازية. ولم تكن هناك إمكانية لتطوير تقنية نووية تعمل للصالح العام. وبعد بضعة أعوامٍ لاحقة، وبعد أن أثبت حدوث التفاعل النووي المتسلسل في مَعمله، كتب سلارد يقول: «لقد أغلقنا كلَّ شيء وتوجَّهنا إلى منازلنا. في تلك الليلة، لم يكن لديَّ شك كبير في أن البؤس سيكون مصير العالم.»

(١-٤) من السابق لأوانه القلق من هذا الأمر

من الشائع رؤية بعض الحكماء وهم يُحاولون تهدئة مخاوف الناس بالإشارة إلى أنه لا يُوجَد ما يمكن القلق بشأنه؛ لأن الذكاء الاصطناعي الذي يُضاهي الذكاء البشري ليس من المُحتمَل أن يظهر قبل عدة عقود. على سبيل المثال، يقول التقرير السابقة الإشارة إليه في القسم السابق إنه «لا تُوجَد أي مَدعاةٍ للقلق من تسبُّب الذكاء الاصطناعي في تهديدٍ وشيك للبشرية».

تلك المُحاجَّة قاصرة في جانبَين. يتمثل الأول في أنها تعد ما يُسمى بمغالطة الرجل القش؛ أي تشويه الحُجة للرد عليها. إن أسباب القلق «لا» تقوم على قُرب حدوث الأمر. على سبيل المثال، كتب نيك بوستروم في كتابه «الذكاء الخارق» يقول: «ليس جزءًا من المُحاجَّة المعروضة في هذا الكتاب القول بأننا على عتبة حدوث إنجازٍ كبير في مجال الذكاء الاصطناعي أو أننا يُمكننا توقُّع، بأي درجةٍ من الدقة، الوقت الذي قد يحدُث فيه هذا التطور». أما الثاني، فيتمثَّل في أن المخاطر الطويلة الأجل يُمكن أن تكون مَدعاةً للقلق الفوري. إن الوقت الصحيح للقلق بشأن تعرُّض البشرية لمشكلةٍ قد تكون خطيرة، يعتمد ليس فقط على وقت حدوث المشكلة، وإنما أيضًا على الوقت الذي سيُستغرَق في وضع حلٍّ لها وتنفيذه.

على سبيل المثال، إذا اكتشفنا أن كُويكبًا كبيرًا في طريقه للاصطدام بالأرض في عام ٢٠٦٩، فهل سنقول إنه لَمن السابق لأوانه القلق بشأن ذلك؟ لا، على العكس تمامًا.

سيكون هناك مشروع طارئ على مستوى العالم لتطوير طريقةٍ لمواجهة هذا التهديد. ولن نتنظِر حتى عام ٢٠٦٨ للبدء في العمل على هذا الحل؛ لأننا لا يُمكننا مُقدمًا تحديد الوقت المطلوب لذلك. في واقع الأمر، مشروع الدفاع الكوكبي التابع لوكالة ناسا يعمل «بالفعل» من أجل التوصُّل إلى حلول مُمكنة لتلك المشكلة، حتى مع العلم بأنه «لا يُوجَد أيُّ احتمالٍ لحدوث اصطدام خطير بالأرض من جانب أي كويكب معروف في المائة عام القادمة». وإن جعلك هذا تشعُر بالرضا، فهم يقولون أيضًا إنه «لم يُكتشَف بعدُ نحو ٧٤ بالمائة من الأجسام القريبة من الأرض، والتي يزيد حجمها عن ٤٦٠ قدمًا».

وإذا نظرنا إلى المخاطر الكارثية العالمية الناتجة عن التغيُّر المناخي، والتي نتوقَّع حدوثها في نهاية هذا القرن، أليس من السابق لأوانه جدًّا التحرُّك لمنعها؟ على العكس، من المُمكن أن نكون قد تأخَّرنا جدًّا. إن الإطار الزمني ذا الصلة لتطوير الذكاء الاصطناعي الخارق يصعُب التنبُّؤ به أكثر، لكن هذا بالطبع يعني أنه، شأنه شأن الانشطار النووي، قد يحدُث أسرع كثيرًا مما توقَّعنا.

إحدى صور مُحاجة «من السابق لأوانه القلق» المعروفة زعم أندرو نج بأن «هذا يُشبه القلق من الزيادة السكانية على كوكب المريخ».[11] (حدَّث نج لاحقًا زعمه بأن استبدل بكوكب المريخ النظام النجمي ألفا سنتوري.) يُعدُّ نج، البروفسير السابق بجامعة ستانفورد، خبيرًا شهيرًا في مجال تعلُّم الآلة، ولآرائه بعض الثقل. إن هذا الزعم يلجأ إلى قياسٍ ملائم: الخطر ليس فقط جرى التعامُل معه بسهولة وإبعاده بعيدًا في المستقبل، وإنما أيضًا من المُستبعَد تمامًا أننا حتى سنحاول نقل مليارات البشر إلى كوكب المريخ في المقام الأول. لكن القياس خاطئ. إننا نُخصِّص «بالفعل» موارد تقنية وعلمية ضخمة لتطوير نظم ذكاء اصطناعي أكثر قوة من ذي قبل، مع عدم توجيه الكثير من الانتباه لما سيحدث إن نجحنا في ذلك. إن القياس الأكثر مُلاءمة، إذن، هو العمل على خطة لنقل الجنس البشري إلى المريخ دون التفكير فيما قد نتنفَّسه أو نشربُه أو نأكله بمجرد وصولنا إلى هناك. قد يصف البعض تلك الخطة بأنها غير حكيمة. أو، يُمكن أن نأخذ كلام نج حرفيًّا ونرى أن إنزال حتى ولو شخص واحد على المريخ سيُعدُّ زيادة سكانية؛ لأن المريخ ليست لديه قدرة استيعابية. ومن ثمَّ فإن المجموعات التي تُخطِّط حاليًّا لإرسال حفنة من البشر إلى المريخ قلقُون بشأن الزيادة السكانية على كوكب المريخ، وهذا هو السبب وراء تطويرهم لنُظُم دعم الحياة.

(١-٥) نحن الخبراء

في كل نقاشٍ حول المخاطر التقنية، يُقدم المعسكر المناهض للتكنولوجيا الزعم القائل بأن كل المخاوف بشأن المخاطر سببها الجهل. على سبيل المثال، يقول أورين إتسيوني، الرئيس التنفيذي لمعهد ألين للذكاء الاصطناعي، والباحث المعروف في مجال تعلُّم الآلة وفهم اللغة الطبيعية:[12]

> عند ظهور أيِّ ابتكارٍ تقني، يُصاب الناس بالخوف. فبدءًا من النسَّاجين الذين كانوا يقذفون أحذيتهم في الأنوال الميكانيكية في بداية الحقبة الصناعية وحتى مخاوف اليوم من ظهور روبوتات قاتلة، استجابتنا يقودها عدم معرفة التأثير الذي ستُحدِثه التقنية الجديدة في إدراكنا لذواتنا ومعيشتنا. وعندما لا نعرف شيئًا، تمدُّنا عقولنا الخائفة بالمعلومات المطلوبة.

نشرت مجلة «بوبيلر ساينس» مقالًا بعنوان «بيل جيتس يخشى الذكاء الاصطناعي، لكن باحثي الذكاء الاصطناعي يعرفون على نحوٍ أفضل» تقول فيه:[13]

> عندما تتحدث إلى الباحثين في مجال الذكاء الاصطناعي — مرةً أخرى، الباحثين الحقيقيين، وهم الأشخاص الذين يتصدَّون لصُنع نُظمٍ عاملة بأيِّ نحوٍ وليس بالطبع عاملة على نحو رائع — تجدهم غير قلقين من احتمالية مُفاجأة الروبوتات ذات الذكاء الخارق لهم، سواء الآن أو في المستقبل. وعلى عكس القصص المُخيفة التي يبدو أن [إيلون] ماسك حريص على سردها، فإن هؤلاء الباحثين لا يبنُون على نحوٍ محموم غرف استدعاء محمية ولا عمليات عدٍّ تنازُلي ذاتية التدمير.

هذا التحليل مُعتمِد على عيِّنة قوامها أربعة من الباحثين، والذين قالوا جميعهم في واقع الأمر في حواراتهم إن الأمان الطويل الأمد للذكاء الاصطناعي كان مسألةً مُهمة.

باستخدام لغة مُماثلة جدًّا للغة المكتوب بها مقال «بوبيلر ساينس»، كتب ديفيد كيني، والذي كان في ذلك الوقت نائب رئيس شركة آي بي إم، خطابًا إلى الكونجرس الأمريكي يتضمَّن الكلمات المُطَمْئنة التالية:[14]

> عندما تستكشف الجوانب العلمية لذكاء الآلات وعندما تُطبِّقها بالفعل في العالم الواقعي للأعمال والمُجتمع، كما فعلنا في شركة آي بي إم لبناء النظام الحاسوبي

المعرفي الرائد الخاص بنا، «واطسون»، تدرك أن تلك التقنية لا تدعم الإشاعات المُقلقة المرتبطة على نحوٍ شائع بالجدل الدائر اليوم حول الذكاء الاصطناعي.

الرسالة واحدة في الحالات الثلاث جميعها: «لا تستمع إليهم؛ فنحن الخبراء». يُمكن الإشارة إلى أن تلك في حقيقة الأمر مُحاجَّة تقوم على القدح الشخصي تُحاول تفنيد الرسالة بالهجوم على أصحابها، لكن حتى لو أخذها المرء على ظاهرها فقط، فإنها واهية. إن إيلون ماسك وستيفين هوكينج وبيل جيتس بالتأكيد على معرفةٍ تامَّة بالتفكير العلمي والتقني، وماسك وجيتس على الخصوص أشرفا على العديد من المشروعات البحثية في مجال الذكاء الاصطناعي واستثمرا فيها. ولن يكون حتى من المعقول القول بأن آلان تورينج وآي جيه جود ونوربرت فينر ومارفن مينيسكي غير مؤهَّلين لمناقشة المسائل المُتعلِّقة بالذكاء الاصطناعي. وأخيرًا، يُشير المقال السابق ذكره والمنشور في مدونة سكوت ألكسندر والذي كان بعنوان «رأي باحثي الذكاء الاصطناعي في مخاطره» إلى أن «باحثي الذكاء الاصطناعي، بما في ذلك بعض القادة في هذا المجال، كان لهم دور مهم في إثارة الانتباه لبعض المسائل المُتعلِّقة بمخاطر الذكاء الاصطناعي والذكاء الخارق منذ البداية». وذكر العديد من هؤلاء الباحثين، والقائمة الآن أطول بكثير.

هناك توجُّه خطابي قياسي آخر من جانب «المُدافِعين عن الذكاء الاصطناعي»، والذي يتمثل في وصف خصومهم بأنهم «لوديون»؛ أي مناهضون للتطوُّر التقني. إن إشارة أورين إتسيوني إلى «النسَّاجين الذين كانوا يقذفون أحذيتهم في الأنوال الميكانيكية» هي ما أقصده هنا؛ إن اللوديين كانوا نسَّاجين حرفيين في أوائل القرن التاسع عشر وكانوا مُعترِضين على إدخال الآلات لتحلَّ محلَّ عملهم اليدوي. وفي عام ٢٠١٥، منحت مؤسسة تكنولوجيا المعلومات والابتكار جائزتها اللودية السنوية لـ «مُثيري الذُّعر فيما يتعلق بدور الذكاء الاصطناعي في نهاية العالم». إنه لتعريف غريب لمصطلح «لودي» أن يتضمَّن تورينج وفينر ومينيسكي وماسك وجيتس، والذين يُعَدون من أبرز المساهمين في التقدُّم التقني الذي حدث في القرنَين العشرين والحادي والعشرين.

إن الاتهام باللودية يعدُّ إساءة فهم لطبيعة المخاوف المثارة والهدف من إثارتها. يبدو الأمر مثل اتهام المهندسين النوويين باللودية إن أشاروا إلى الحاجة للتحكُّم في التفاعلات الانشطارية. وكما هو الحال مع الظاهرة الغريبة المُتمثِّلة في الزعم المفاجئ من جانب باحثي الذكاء الاصطناعي بأن الذكاء الاصطناعي مُستحيل، أعتقد أننا يُمكننا إرجاع هذا التوجُّه المُحيِّر إلى القبلية التي تحاول الدفاع عن التقدم التقني.

(٢) التهرُّب

بعض المعلقين مُستعدُّون للإقرار بأن المخاطر حقيقية، لكنهم يقدمون حججًا ترى ضرورة عدم فعل أي شيء. وتتضمن تلك الحجج استحالة فعل أي شيءٍ، وأهمية فعل شيء آخر تمامًا، والحاجة للسكوت عن المخاطر.

(٢-١) عدم إمكانية التحكُّم في الأبحاث

من الردود الشائعة على الآراء التي ترى أن الذكاء الاصطناعي المُتطور قد يُعرِّض البشر لمخاطر، الزعم بأن إيقاف أبحاث الذكاء الاصطناعي مُستحيل. لاحظ القفزة العقلية هنا: «حسنًا، إن شخصًا ما يناقش المخاطر! لا بدَّ أنه يقترح وقف بحثي!» قد تكون تلك القفزة العقلية مُلائمة عند مناقشة المخاطر اعتمادًا فقط على مشكلة الغوريلا، وأنا أميل إلى الموافقة على أن حلَّ مشكلة الغوريلا بمنع بناء الذكاء الاصطناعي الخارق سيتطلَّب وضع نوعٍ من القيود على الأبحاث في مجال الذكاء الاصطناعي.

لكن النقاشات الحديثة التي دارت حول المخاطر ركَّزت ليس على مشكلة الغوريلا العامة (باللغة الصحفية، النزال الشديد بين البشر والذكاء الخارق)، ولكن على مشكلة الملك ميداس وصورها المختلفة. إن حلَّ مشكلة الملك ميداس يحلُّ أيضًا مشكلة الغوريلا؛ ليس عن طريق منع بناء الذكاء الاصطناعي الخارق أو إيجاد طريقة للتغلب عليه، وإنما بضمان عدم دخوله على الإطلاق في صراعٍ مع البشر في المقام الأول. إن النقاشات الدائرة حول مشكلة الملك ميداس بوجهٍ عام تتجنَّب اقتراح ضرورة تقييد البحث في مجال الذكاء الاصطناعي؛ فهي تقترح فقط أنه يجب الاهتمام بمسألة منع المخاطر التي قد تنتج عن النظم السيئة التصميم. في نفس الإطار، إن مناقشة مخاطر التسرُّب في المحطات النووية يجب تفسيرها ليس باعتبارها محاولةً لوقف الأبحاث في مجال الفيزياء النووية وإنما كإشارةٍ لضرورة توجيه مزيدٍ من الجهود على حلِّ مُشكلة التسرُّب.

هناك، بالمُصادفة، سابقة تاريخية مُثيرة جدًّا للاهتمام فيما يتعلق بإيقاف الأبحاث. ففي أوائل سبعينيات القرن الماضي، بدأ علماء البيولوجيا القلق من أن طُرق الحمض النووي التركيبي الحديثة — والتي يحدث فيها نقل جينات من كائنٍ لآخر — قد تؤدي إلى مخاطر كبيرة على صحة الإنسان والنظام البيئي العالمي. أدَّى مؤتمران في منتجع أسيلومار في كاليفورنيا في عامي ١٩٧٣ و١٩٧٥ أولًا إلى تعليق هذه التجارب ثم إلى

توجيهات مُفصلة تتعلق بالأمان البيولوجي تتلاءم مع المخاطر التي تفرضها أي تجربةٍ مقترحة.[15] بعض فئات هذه التجارب، مثل تلك التي كانت تتضمَّن جينات سموم، عُدت خطيرة جدًّا بحيث لم يعُد في الإمكان إتاحة إجرائها.

بعد المؤتمر الذي عقد في عام ١٩٧٥ مباشرة، بدأت المعاهد الوطنية للصحة، التي تمول تقريبًا كل الأبحاث الطبية الأساسية في الولايات المتحدة، عملية إنشاء اللجنة الاستشارية الخاصة بالحمض النووي التركيبي. كان لتلك اللجنة دور مُهم في تطوير توجيهات المعاهد الوطنية للصحة التي نفَّذت بالأساس توصيات مؤتمر أسيلومار. ومنذ عام ٢٠٠٠، تضمَّنت تلك التوجيهات منع الموافقة على تمويل أي بروتوكول يتضمَّن «تغيير الجينوم البشري»؛ أي تعديل الجينوم البشري بطُرُق يمكن توريثها للأجيال القادمة. وهذا المنع تبعه حظر قانوني في أكثر من خمسين دولة.

إن هدف «تحسين السلالة البشرية» كان أحد أحلام حركة تحسين النسل في أواخر القرن التاسع عشر وأوائل القرن العشرين. وقد أعاد إحياء هذا الحلم تطوير «كريسبر-كاس٩»، وهي طريقة دقيقة جدًّا لتعديل الجينوم. لقد ترك مؤتمر دولي عُقد في عام ٢٠١٥ الباب مفتوحًا أمام التطبيقات المستقبلية، داعيًا إلى وضع قيود حتى «وجود إجماع مجتمعي واسع حول مدى مُلاءمة التطبيق المُقترح».[16] وفي نوفمبر من عام ٢٠١٨، أعلن العالم الصيني خه جيان كوي تعديله لجينومات ثلاثة أجنَّة بشرية، اثنان منهم على الأقل اكتمل نموُّهما وأصبحا طفلَين. وتبع ذلك اعتراضات دولية قوية، وفي وقت كتابة هذا الكتاب، يبدو أن جيان كوي قيد الإقامة الجبرية في منزله. وفي مارس ٢٠١٩، طالبت هيئة دولية من كبار العلماء صراحةً بتعليقٍ رسميٍّ لتلك التجارب.[17]

إن الدرس المستفاد من هذا الجدل فيما يتعلق بالذكاء الاصطناعي مُزدوج. فمن جانب، إنه يوضح أننا «يمكننا» وقف العمل في أيِّ مجالٍ بحثيٍّ له مخاطر ضخمة. إن الإجماع الدولي على حظر تعديل الجينوم ناجح على نحوٍ شبه كامل حتى الآن. ولم يتحقَّق الخوف من أن الحظر سيؤدي إلى القيام بالأبحاث في الخفاء أو في دول لا تُعارض ذلك. ومن جانب آخر، تعديل الجينوم عملية يسهل التعرف عليها، وهي حالة استخدام محدَّدة لمعرفةٍ عامةٍ أكثر، خاصة بعلم الوراثة، تتطلب معدات خاصة وبشرًا لإجراء التجارب عليهم. علاوة على ذلك، إنها عملية تتبع مجالًا — وهو الطب التناسُلي — خاضعًا بالفعل لمُراقبةٍ دقيقة وتشريعات صارمة. وتلك السمات لا تنطبق على الذكاء الاصطناعي العام، وحتى الآن، لم يخرج علينا أيُّ أحدٍ بأي صيغةٍ معقولة يمكن لأي تشريع، لتقييد البحث في مجال الذكاء الاصطناعي، أن يتخذها.

(٢-٢) الماذاعنية

لقد تعرَّفتُ على مصطلح «الماذاعنية» على يد مستشارٍ لسياسي بريطاني كان عليه التعامل معه على نحوٍ منتظم في اللقاءات العامة. فبصرف النظر عن موضوع الكلمة التي كان يُلقيها، كان شخص يسأله على نحو دائم: «ماذا عن القضية الفلسطينية؟»

ردًّا على أي ذِكر لمخاطر الذكاء الاصطناعي المُتطور، من المُحتمَل أن يستمع المرء إلى السؤال الآتي: «ماذا عن فوائد الذكاء الاصطناعي؟» على سبيل المثال، يقول أورين إتسيوني:[18]

> التوقعات التشاؤمية عادة ما تفشل في أن تأخذ في الاعتبار المزايا المُحتملة للذكاء الاصطناعي والمتعلقة بمنع الأخطاء الطبية وتقليل حوادث المركبات وغير ذلك.

وها هو مارك زوكربيرج، الرئيس التنفيذي لشركة فيسبوك، يقول في حوار حديث تداولته وسائل الإعلام مع إيلون ماسك:[19]

> إذا كنت تعارض الذكاء الاصطناعي، فأنت إذن تُعارض السيارات الأكثر أمانًا التي لن تتعرض لحوادث، وتُعارض التشخيص الأفضل للناس عندما يمرضون.

إذا نحَّينا جانبًا المفهوم القبلي الذي يرى أن أي شخصٍ يذكر المخاطر «يُعارض الذكاء الاصطناعي»، فإن كلًّا من زوكربيرج وإتسيوني يريان أن الحديث عن المخاطر يعني تجاهل المزايا المُحتملة للذكاء الاصطناعي أو حتى إنكار وجودها.

هذا بالطبع نوع من الغباء، لسببَين. أولًا: إذا لم تكن هناك فوائد مُحتملة للذكاء الاصطناعي، فلن يكون هناك أيُّ دافع اقتصادي أو اجتماعي للقيام بأبحاثٍ في مجال الذكاء الاصطناعي؛ ومن ثمَّ لن يُوجَد أبدًا خطر الوصول إلى الذكاء الاصطناعي الذي يُضاهي الذكاء البشري. إننا ببساطة حينها لن نحتاج إلى هذا الجدل على الإطلاق. ثانيًا: «إن لم يجر الحدُّ من المخاطر بنجاح، فلن تكون هناك فوائد». إن الفوائد المُحتملة للطاقة النووية قلَّت على نحوٍ كبير بسبب انصهار قلب المُفاعل الجزئي في محطة ثري مايل آيلاند في عام ١٩٧٩، والانبعاثات الكارثية والتفاعلات النووية غير المُسيطَر عليها في تشيرنوبل في عام ١٩٨٦ والانصهارات المُتعدِّدة التي حدثت في فوكوشيما في عام ٢٠١١. لقد حدَّت تلك الكوارث من نموِّ الصناعة النووية. فقد هجرت إيطاليا الطاقة النووية في عام ١٩٩٠،

وأعلنت كلٌّ من بلجيكا وألمانيا وإسبانيا وسويسرا عن نيتها فعل ذلك. ومنذ عام ١٩٩٠، بلغ المعدل العالمي لإنشاء المحطات النووية نحو عُشر ما كان عليه قبل كارثة تشيرنوبل.

(٢-٣) السكوت

إن أكثر أشكال التهرب تطرُّفًا هو ببساطةٍ القولُ بأننا يجب أن نسكُت عن مسألة المخاطر. على سبيل المثال، تقرير «دراسة المائة عام حول الذكاء الاصطناعي» السابق الإشارة إليه يتضمَّن التحذير التالي:

> إذا تعامل المجتمع مع هذه التقنيات على نحوٍ أساسي بخوف وشك، فستحدث عثرات من شأنها أن تُبطئ من تطوُّر الذكاء الاصطناعي أو تجعله يتمُّ في الخفاء، مما يعوق القيام بالعمل المهم المُتعلق بضمان أمان واعتمادية تقنيات الذكاء الاصطناعي.

قدم روبرت أتكينسون، رئيس مؤسسة تكنولوجيا المعلومات والابتكار (نفس المؤسسة التي تمنح الجائزة اللودية) مُحاجَّة مُماثلة في نقاشٍ جرى في عام ٢٠١٥.[20] فبينما هناك اعتراضات وجيهة حول الطريقة التي يجب من خلالها وصف المخاطر عند التحدُّث إلى وسائل الإعلام، فإن الرسالة الإجمالية واضحة: «لا تذكر المخاطر؛ إذ سيؤثر ذلك على مسألة التمويل». بالطبع، إن لم يعلم أحد بوجود مخاطر، فلن يكون هناك تمويل للأبحاث المتعلِّقة بالحد من المخاطر ولا سبب يدعو أحدًا للعمل عليها.

قدم عالم النفس المعرفي المعروف ستيفين بينكر صورةً أكثر تفاؤلًا من مُحاجة أتكينسون. ففي رأيه أن «ثقافة الأمان في المجتمعات المتقدمة» ستضمن الحدَّ من كل المخاطر المهمة للذكاء الاصطناعي؛ ومن ثمَّ فمن غير المُلائم ومن غير المفيد لفت الأنظار إلى تلك المخاطر.[21] حتى إذا غضضْنا الطرف عن حقيقة أن ثقافة الأمان المُتقدِّمة الخاصَّة بنا قد أدَّت إلى كارثتي تشيرنوبل وفوكوشيما والاحتباس الحراري الجامح، فإن مُحاجَّة بينكر قد جانبها الصواب تمامًا. إن ثقافة الأمان تقوم بالأساس على أشخاص يلفتُون الأنظار إلى أنماط الفشل المُمكنة ويجدُون سُبلًا لضمان عدم حدوثها. (وفيما يتعلق بالذكاء الاصطناعي، النموذج القياسي هو نمط الفشل.) والقول بأنه من السخيف لفت الانتباه إلى أيِّ نمط فشل لأنَّ ثقافة الأمان ستتعامل مع ذلك على أيِّ حالٍ يُشبه القول بأن

لا أحد يجب أن يستدعي سيارة إسعاف عندما يرى حادث سير هرب فيه السائق وترك الشخص المُصاب في الشارع لأنَّ شخصًا ما سيستدعيها.

عند مُحاولة تعريف العامة وصُنَّاع السياسات بالمخاطر، يكون باحثو الذكاء الاصطناعي في وضعٍ أسوأ مقارنةً بالفيزيائيِّين النوويين. فهؤلاء الفيزيائيون لا يحتاجون إلى تأليف كتبٍ تُوضِّح للعامة أن تَجمُّع كتلةٍ حرجة من اليورانيوم العالي التخصيب قد يُمثِّل خطرًا؛ لأن عواقب ذلك قد تجسَّدت بالفعل في هيروشيما وناجاساكي. فلا يحتاج الأمر إلى مزيدٍ من الجهد لإقناع الحكومات والجهات التمويلية بأن عامل الأمان مُهم في تطوير الطاقة النووية.

(٣) القبلية

في رواية باتلر «إريون»، أدَّى التركيز على مشكلة الغوريلا إلى انقسام سابق لأوانه وخاطئ بين مؤيدي الآلات ومعارضيها. يعتقد مؤيدو الآلات أن خطر هيمنة الآلات محدود أو غير موجود؛ في حين يعتقد مُعارضوها أنه لا يمكن مواجهته ما لم تُدمَّر كل الآلات. يُصبح الصراع قبليًّا، ولا يُحاول أحد حل المشكلة الأساسية المتمثلة في إبقاء البشر الآلات تحت سيطرتهم.

بدرجات مختلفة، تخضع كل المسائل التقنية المهمة في القرن العشرين — الطاقة النووية والكائنات المعدلة وراثيًّا وأنواع الوقود الحفري — للقبلية. في كل مسألة، هناك جانبان، جانب مؤيد وجانب معارض. إن ديناميكيات ونواتج كل منهما كانت مختلفة، لكن أعراض القبلية واحدة: تشويه السمعة وعدم الثقة المتبادَلَين والحجج غير العقلانية ورفض قبول أي نقطة (منطقية) قد تكون في صالح الجانب الآخر. فيسعى الجانب المؤيد للتقنية لإنكار وإخفاء المخاطر، ويصاحب ذلك اتهامات باللودية؛ في حين يرى الجانب المعارض للتقنية أن المخاطر لا يمكن مواجهتها وأن المشكلات غير قابلة للحل. إن أي عضوٍ في الجانب المؤيد للتقنية والذي يكون أمينًا للغاية فيما يتعلق بمشكلةٍ ما يُرى على أنه خائن، وهو أمر مُحزن بوجهٍ خاص؛ نظرًا لأن الجانب المؤيد للتقنية عادة ما يتضمَّن معظم الناس المؤهلَين لحل المشكلة. كما أن عضو الجانب المعارض للتقنية الذي يناقش الحلول المحتملة خائن هو الآخر لأنه يرى أن التقنية نفسها هي مصدر الشر وليس آثارها المحتملة. وبهذه الطريقة، يمكن فقط للأصوات الأكثر تطرفًا — تلك التي يقلُّ بشدة احتمال الاستماع إليها من قبل الجانب الآخر — أن تتحدَّث باسم كل جانب.

في عام ٢٠١٦، دُعيت إلى الذهاب إلى مقر رئيس وزراء بريطانيا للقاء بعض مستشاري ديفيد كاميرون الذي كان رئيس الوزراء حينها. كان هؤلاء المستشارون قلقين من أن الجدل الدائر حول الذكاء الاصطناعي كان على وشك أن يُشبه الجدل الدائر حول الكائنات المُعدَّلة وراثيًّا، الذي أدى في أوروبا إلى ما اعتبره المستشارون تشريعاتٍ سابقة لأوانها ومُقيدة للغاية فيما يتعلق بإنتاج تلك الكائنات وتسميتها. وأرادُوا تجنُّب حدوث نفس الشيء فيما يتعلَّق بالذكاء الاصطناعي. إن لمخاوفهم بعض الوجاهة؛ فالجدل حول الذكاء الاصطناعي أصبح في خطر التحول إلى جدل قبلي؛ أي في تكوين معسكرَين أحدهما مؤيد للذكاء الاصطناعي والآخر معارض له. وسيضر هذا بالمجال لأنه ببساطة من الخطأ أن يعدَّ القلق بشأن المخاطر المتضمنة في الذكاء الاصطناعي المتطور موقفًا مُعاديًا للذكاء الاصطناعي. فالفيزيائي القلق بشأن مخاطر الحرب النووية أو خطر انفجار مفاعل نووي سيئ التصميم ليس «مُعاديًا للفيزياء». والقول بأنَّ الذكاء الاصطناعي سيكون قويًّا بالقدر الكافي بحيث يكون له تأثير عالمي ثناءٌ على المجال وليس هجومًا عليه.

من المهم أن يعترف مجتمع الذكاء الاصطناعي بوجود مخاطر ويعمل على الحدِّ منها. إن المخاطر، إلى الحدِّ الذي نفهمُه عنها، ليست محدودةً ولا صعبًا منعُها. نحن نحتاج لبذل قدْر كبير من الجهد لتجنُّبها، بما في ذلك إعادة تشكيل وإعادة بناء أسس الذكاء الاصطناعي.

(٤) ألا يمكننا فقط أن ...؟

(٤-١) ... نوقف تشغيل الآلات؟

إن الكثير من الناس، بما فيهم أنا، بمجرد أن يفهموا الفكرة الأساسية للخطر الوجودي — سواء في شكل مُشكلة الغوريلا أو مشكلة الملك ميداس — سيبدءون على الفور في البحث عن حلٍّ سهل. في الغالب، أول شيء سيطرأ على ذهنهم هو إيقاف تشغيل الآلات. على سبيل المثال، آلان تورينج نفسه، كما ذكرنا آنفًا، يتكهَّن بأننا يُمكن أن «نُبقي الآلات في وضعٍ تابع لنا، على سبيل المثال بإيقاف تشغيلها في اللحظات الحاسمة».

هذا لن يجدي، للسبب البسيط المُتمثِّل في أن الكيان الخارق الذكاء سوف «يفكر في تلك الاحتمالية»، ويتَّخذ خطواتٍ لمنعها. وسيفعل هذا ليس لأنه يُريد البقاء والاستمرار، ولكن لأنه يسعى إلى تحقيق الهدف الذي ندمجه فيه ويعرف أنه إن فشل في ذلك فسيُوقف تشغيله.

هناك بعض النُّظم التي خضعت للدراسة والتي في الواقع لا يُمكن إيقاف تشغيلها دون تدمير جانب كبير من ثمار حضارتنا. إنها النُّظُم المُنفذة على هيئة ما يُسمى بالعقود الذكية في سلسلة الكتل (البلوك تشين). إن «سلسلة الكتل» تُتيح التوزيع الواسع النطاق للقُدرة الحوسبية وحفظ السجلات اعتمادًا على التشفير؛ إنها مُصمَّمة بوجهٍ خاص بحيث لا يُمكن حذف أي عنصر بيانات أو إيقاف تنفيذ أي عقدٍ ذكي دون التحكم بالأساس في عددٍ كبير للغاية من الآلات وإلغاء السلسلة، مما قد يؤدِّي بالتبعية إلى تدمير جزءٍ كبير من الإنترنت و/أو النظام المالي. إنه لمحلُّ نزاعٍ التساؤل ما إذا كانت البراعة غير العادية سمةً مُبتكرة أم عيبًا. إنها بالتأكيد أداة يُمكن أن يستخدمها أي نظام ذكاء اصطناعي خارق لحماية نفسه.

(٤-٢) ... نقيد الآلات؟

إذا لم يكن بإمكاننا إيقاف تشغيل نظم الذكاء الاصطناعي، فهل يُمكنُنا تقييد الآلات بنظام حمايةٍ من نوعٍ ما، بحيث نحصُل منها على إجابات مُفيدة على الأسئلة لكن دون أن نسمح لها بالتأثير على العالم الواقعي على نحوٍ مُباشر؟ تلك هي الفكرة وراء نُظُم الذكاء الاصطناعي الخاصة بشركة أوراكل، والتي جرت مناقشتها على نحوٍ مُطوَّل في أوساط المُهتمين بمسألة أمان الذكاء الاصطناعي.[22] إن أي نظام ذكاء اصطناعي خاص بأوراكل يمكن أن يكون ذكيًّا على نحوٍ عشوائي، لكن تُمكنه الإجابة فقط بنعم أو لا (أو إعطاء أي احتمالاتٍ مشابهة) فيما يتعلق بأي سؤال. ويمكنه الوصول إلى كل المعلومات التي يمتلكها الجنس البشري من خلال اتصال للقراءة فقط؛ أي إنه ليس لديه اتصال مباشر بالإنترنت. بالطبع، هذا يعني وقف محاولة تطوير الروبوتات والمساعدين الخارقين والعديد من الأنواع الأخرى من نُظُم الذكاء الاصطناعي، لكن أي نظام ذكاء اصطناعي جيِّدٍ خاصٍّ بأوراكل ستكون له قيمة اقتصادية هائلة؛ لأننا يُمكننا أن نسأله أسئلةً إجابتها مُهمَّة لنا؛ مثل ما إذا كان مرض ألزهايمر سببه كائن مُعدٍ أو ما إذا كان من الجيد حظر الأسلحة الذاتية التشغيل. لذا، فإن نظم الذكاء الاصطناعي الخاصة بأوراكل بالتأكيد تُعَدُّ خيارًا مثيرًا للاهتمام.

لسوء الحظ، تُوجَد بعض الصعوبات المهمَّة. أولًا: إن نظام الذكاء الاصطناعي الخاص بأوراكل سيكون على الأقل مُفيدًا في فهم الجوانب الفيزيائية لعالَمه وأصوله — الموارد

الحاسوبية ونمط تشغيلها والكيانات الغامضة التي شكلت مخزونه من المعلومات وتطرح عليه أسئلة الآن — كما هو الحال بالنسبة لنا فيما يتعلق بفهم عالمنا. ثانيًا: إذا كان الهدف من نظام الذكاء الاصطناعي الخاص بأوراكل هو تقديم إجاباتٍ دقيقة على الأسئلة في قدرٍ معقُول من الوقت، فسيكون لديه الدافع للخروج من القفص الذي هو فيه لاكتساب المزيد من الموارد الحاسُوبية والتحكُّم في مُوجِّهي الأسئلة إليه بحيث يسألونه فقط أسئلةً بسيطة. وأخيرًا، نحن لم نبتكِر بعدُ نظام حماية آمنًا ضد تدخُّل البشر العاديين، فضلًا عن الآلات الخارقة.

أعتقد أنه «ربما» تكون هناك حلول لبعض هذه المُشكلات، وخاصةً إذا قيَّدنا نُظُم الذكاء الاصطناعي الخاصة بأوراكل بحيث تكون آلات حاسبة منطقية أو بايزية جيِّدة على نحوٍ واضح. هذا يعني أننا يمكن أن نُصرَّ على أن الخوارزمية يمكن أن تنتج فقط نتيجةً تُتيحها المعلومات المُعطاة، ويمكننا التحقُّق رياضيًّا من تحقيق الخوارزمية لهذا الشرط. لكن هذا لا يحلُّ مشكلة التحكُّم في العملية التي ستُحدِّد «أي» العمليات الحسابية المنطقية أو البايزية التي سيتمُّ إجراؤها، من أجل الوصول إلى أقوى نتيجةٍ مُحتملة بأسرع ما يُمكن. ولأن تلك العملية لديها دافع للتفكير بسرعة، فإن لديها دافعًا لاكتساب موارد حاسوبية بالطبع للحفاظ على وجودها.

في عام ٢٠١٨، عقد مركز الذكاء الاصطناعي المُتوافق مع البشر التابع لجامعة كاليفورنيا ببيركلي ورشة عمل طرحنا فيها السؤال الآتي: «ماذا ستفعل إذا كنتَ تعرف على وجه اليقين أن الذكاء الاصطناعي الخارق سيتحقَّق في غضون عقد؟» كانت إجابتي هي إقناع المُطوِّرين أن يُؤجِّلوا بناء الكيان الذكي العام — ذلك الذي يُمكنه اختيار أفعاله في العالم الواقعي — ويبنُوا بدلًا منه نظام ذكاء اصطناعي خاصًّا بأوراكل. وفي تلك الأثناء، سنعمل على حلِّ المُشكلة بجعل نظم الذكاء الاصطناعي الخاصة بأوراكل آمنةً على نحوٍ مُثبت إلى أقصى حدٍّ مُمكن. وإمكانية نجاح تلك الاستراتيجية ترجع إلى أمرَين؛ أولًا: ستبلُغ القيمة المالية لنظام الذكاء الاصطناعي الخارق الخاص بأوراكل تريليونات الدولارات، مما قد يجعلُ المُطوِّرين على استعدادٍ لقبول هذا القيد؛ وثانيًا: التحكم في نظم الذكاء الاصطناعي الخاصة بأوراكل أسهل على نحوٍ شبه مؤكَّد من التحكُّم في كيانٍ ذكي عام، لذا، ستكون لدَينا فرصة أفضل لحل المشكلة خلال العقد.

(٤-٣) ... نعمل في فِرَقٍ يتعاون فيها البشر والآلات؟

هناك فكرة منتشرة في عالم الأعمال مفادها أن الذكاء الاصطناعي لن يُمثل أيَّ تهديد على العمالة أو على البشرية؛ لأننا حينها ستكون لدَينا فِرَق تعاونية مكونة من البشر والآلات. على سبيل المثال، ذكر الخطاب الذي وجهه ديفيد كيني للكونجرس الأمريكي، والذي عرضنا له قبل ذلك في هذا الفصل، أن «نُظُم الذكاء الاصطناعي العالية القيمة مُصمَّمة على نحوٍ خاص لكي تدعم الذكاء البشري، وليس لكي تحلَّ محلَّ العُمال».[23]

في حين أن أحد المُتهكِّمين قد يُشير إلى أن هذه مجرد خدعة دعائية لتيسير عملية حذف الموظفين البشريِّين من قوائم عملاء الشركات، فأنا أعتقد أن هذا يُحرك الأمر إلى الأمام قليلًا. إن الفِرَق التعاونية المُكوَّنة من البشر والآلات الذكية لهي في واقع الأمر هدف مرغوب فيه. لكن من المعروف أن أي فريق لن يكون ناجحًا إلا إذا كانت أهداف أعضائه مُتوافقة، لذا، فإن التأكيد على الفِرَق التعاونية المُكوَّنة من البشر والآلات الذكية يُبرز الحاجة إلى حلِّ المشكلة الأساسية المتعلقة بتوفيق القيمة. وبالطبع، إبراز المشكلة يختلف عن حلها.

(٤-٤) ... نندمج مع الآلات؟

عندما تتطوَّر عملية تكوين فرق تعاونية من البشر والآلات إلى أقصى حد، تُصبح عملية دمج بين الإنسان والآلة تلحق فيها مكوِّنات إلكترونية مباشرة بالدماغ وتُشكل جزءًا من كيانٍ واحدٍ ومُمتدٍّ وواعٍ. يصف عالم المُستقبليات راي كيرزويل تلك الاحتمالية كما يلي:[24]

> نحن سنندمج معها مباشرة، وسنُصبح آلات ذكية. ... وعندما نصلُ إلى أواخر ثلاثينيات أو أربعينيات القرن الحادي والعشرين، سيُصبح تفكيرنا غير بيولوجي على نحوٍ غالب، والجزء غير البيولوجي سيكون في النهاية ذكيًّا للغاية وستكون لديه قدرات عالية بحيث يُمكنه نمذجة ومُحاكاة وفهم الجزء البيولوجي على نحوٍ كامل.

يرى كيرزويل تلك التطوُّرات على نحوٍ إيجابي. أما إيلون ماسك، على الجانب الآخر، فيرى أن عملية الدمج بين البشر والآلات بالأساس استراتيجية دفاعية:[25]

> إن حقَّقنا تكافُلًا كاملًا، فلن تكون الآلة الذكية «كيانًا آخر»؛ ستكون هي أنت و[ستكون لها] علاقة بالقشرة الدماغية الخاصة بك تُماثل علاقة قشرتك

> الدماغية بنظامك الطرفي. ... سيكون لدَينا الاختيار ما بين أن يجري تجاهُلنا ونُصبح فعليًّا بلا فائدة أو أشبه بحيوان أليف — مثل قطٍّ منزلي أو ما شابه — أو نصِل في النهاية إلى طريقةٍ يُمكن من خلالها أن نتكافل مع الآلات الذكية ونندمج معها.

إن شركة نيورالينك التي أسسها ماسك تعمل على تطوير جهازٍ يُسمَّى «الرابط العصبي» والذي يقوم على تقنيةٍ جرى وصفها في سلسلة روايات «الثقافة» التي كتبها إين بانكس. إن الهدف هو الربط على نحوٍ فعَّال ودائم بين القشرة الدماغية والشبكات والنظم الحاسوبية الخارجية. لكن هناك عقبتان فنِّيتان أساسيتان؛ أولًا: صعوبات الربط بين الجهاز الإلكتروني والنسيج البشري، وإمداده بالطاقة، وربطه بالعالَم الخارجي؛ وثانيًا: حقيقة أننا لا نفهم تقريبًا شيئًا عن التنفيذ العصبي للمستويات الأعلى من المعرفة في الدماغ، لذا، نحن لا نعرف أين سنربط الجهاز وعمليات المعالجة التي يجب أن يقوم بها.

أنا غير مُقتنع تمامًا بأن العقبتَين المذكورتَين في الفقرة السابقة لا يُمكن تجاوزهما. أولًا: تقنيات مثل «الغبار العصبي» تُقلل على نحوٍ سريع من مُتطلبات الحجم والطاقة الخاصة بالأجهزة الإلكترونية التي يُمكن إرفاقها بالعصبونات وتُوفِّر عمليات استشعار ومُحاكاة واتصال عبر الجمجمة.[26] (التقنية وصولًا إلى عام ٢٠١٨ قد وصلت إلى حجم يصلُ إلى نحو ملِّيمتر مُكعب واحد، لذا، فإن «الحُبَيبة العصبية» قد تكون مُصطلحًا أكثر دقة.) ثانيًا: الدماغ نفسه يمتلك قدرات هائلة على التكيُّف. كان يعتقد، على سبيل المثال، أننا سيكون علينا فهم الشفرة التي يستخدمها الدماغ للتحكم في عضلات الذراع قبل أن يكون بإمكاننا بنجاح توصيل أحد الأدمغة بذراعٍ آليةٍ، وأننا سيكون علينا فهم طريقة تحليل قوقعة الأذن للصوت قبل أن يُمكننا زرع جزءٍ بديل لها. واتضح، بدلًا من ذلك، أن الدماغ يقوم بمعظم العمل بالنيابة عنا. فهو يتعلم بسرعة كيف يجعل ذراع الروبوت تفعل ما يُريده مالكها، وكيف يُحول نتاج قوقعة الأذن البديلة إلى أصوات مفهومة. من الممكن تمامًا أن نتوصَّل إلى طُرق لتزويد الدماغ بذاكرةٍ إضافية وبقنوات اتصال مع أجهزة الكمبيوتر وربما حتى بقنوات اتصال مع أدمغةٍ أخرى؛ كل ذلك دون حتى أن نفهم فعليًّا كيف يعمل أيٌّ منها.[27]

بصرف النظر عن مدى الجدوى التقنية لتلك الأفكار، يجب على المرء أن يسأل إن كان هذا التوجُّه يُعدُّ أفضل مستقبل ممكن للبشرية أم لا. إن احتاج البشر إلى جراحة في

الدماغ فقط لمواجهة التهديد الذي تفرضه التقنيات التي ابتكروها، فربما ارتكبنا خطأً ما في موضعٍ ما أثناء العملية.

(٤-٥) ... نتجنَّب دمج أهداف بشرية؟

أحد الاعتقادات الشائعة يرى أن سلوكيات الذكاء الاصطناعي المُسبِّبة للمُشكلات تنبع من دمج «أنواع» مُعينة من الأهداف؛ فإن جرى تجنُّب هذا، فسيكون كل شيء على ما يرام. لذا، على سبيل المثال، إن يان ليكن، الذي يُعَد أحد رواد مجال التعليم المُتعمِّق والذي يترأس قسم الأبحاث الخاصة بالذكاء الاصطناعي في شركة فيسبوك، عادةً ما يذكُر تلك الفكرة عندما يُقلل من شأن الخطر الذي قد ينجم عن الذكاء الاصطناعي:[28]

> لا يُوجد سبب لدى الآلات الذكية لامتلاك غرائز خاصة بالحفاظ على الذات أو غيرة أو ما إلى ذلك. ... فهي لن تكون لديها تلك «العواطف» المُدمِّرة ما لم ندمجها فيها. وأنا لا أفهم سبب رغبتنا في فعل ذلك.

في نفس الإطار، يُوفِّر ستيفين بينكر تحليلًا يقوم على الجنس:[29]

> يُسقط أصحاب السيناريوهات التشاؤمية فيما يتعلق باستخدام الذكاء الاصطناعي فكرة الذَّكَر المُهيمن الضيقة الأفق على مفهوم الذكاء. فهم يفترضُون أن الروبوتات الذكية إلى حدٍّ خارق ستُطور أهدافًا مثل التخلُّص من أسيادها أو السيطرة على العالم. ... من الغريب أن الكثير من هؤلاء لا يرَون احتمالية أن الذكاء الاصطناعي سيتطور طبيعيًّا على نحوٍ أُنثوي؛ إذ سيكون قادرًا على نحوٍ كامل على حلِّ المشكلات، لكن دون أن تكون لدَيه أي رغبة في قتل الأبرياء أو السيطرة على الحضارة.

كما أوضحنا من قبل في النقاش الخاص بالأهداف الأداتية، لا يهم ما إذا كنا ندمج «عواطف» أو «رغبات» مثل الحفاظ على الذات أو امتلاك الموارد أو اكتشاف المعرفة أو، في الحالة المتطرفة، السيطرة على العالم. إن الآلة ستمتلك تلك العواطف على أيِّ حال، باعتبارها أهدافًا فرعية لأي هدفٍ ندمجه فيها، وبصرف النظر عن جنسها. فبالنسبة

إلى أي آلة، الموت ليس سيئًا في حدِّ ذاته. لكن يجب تجنُّبه لأنه من الصعب جلب فنجان القهوة إذا كنت ميتًا.

هناك حلٌّ أكثر تطرُّفًا؛ وهو تجنُّب دمج أي أهداف في الآلة. ربما تقول إن المشكلة تكون هكذا قد حُلت. للأسف، الأمر ليس بسيطًا على هذا النحو. فبدون أهداف، لا يُوجَد أي ذكاء: لا يُوجَد اختلاف بين أي فعل والآخر، وستكون الآلة أشبه بمُولِّد أعداد عشوائية. وبدون أهداف، فلن يكون أيضًا هناك سبب للآلة لتفضيل جنَّة البشر على كوكبٍ تحوَّل إلى بحر من مشابك الورق (وهو سيناريو وُصف على نحوٍ مُفصَّل من قِبَل نيك بوستروم). في الواقع، قد تكون النتيجة الأخيرة مثاليةً لبكتيريا ثايوباسيلس فيراكسدانز الآكلة للحديد. ففي غياب مفهومٍ ما عن أهمية التفضيلات البشرية، على أي أساس يمكن القول بأن تلك البكتيريا على خطأ؟

يُوجَد شكل مختلف من فكرة «تجنُّب دمج الأهداف» والذي يتمثَّل في مفهوم أن النظام الذكي بالقدر الكافي بالضرورة سيُطوِّر، نتيجةً لذكائه، الأهداف «الصحيحة» من تلقاء نفسه. في الغالب، مُناصرو هذا المفهوم مُعجبون بالنظرية القائلة بأن الأشخاص ذوي الذكاء الأكبر يميلون إلى أن تكون لديهم أهداف غيرية ونبيلة أكثر؛ وهو رأيٌ قد يكون مُرتبطًا بمفهوم المؤيدين لذواتهم.

إن فكرة أنه من المُمكن إدراك الأهداف في العالم قد نُوقشت باستفاضةٍ من قبل فيلسوف القرن الثامن عشر الشهير ديفيد هيوم في عمله «رسالة في الطبيعة البشرية».[30] لقد سمَّاها «مشكلة ما يجب أن يكون» وخلص إلى أنه ببساطة، من الخطأ الاعتقاد أن الواجبات الأخلاقية يمكن استخلاصها من الحقائق الطبيعية. لمعرفة السبب، انظر، على سبيل المثال، في تصميم لوح الشطرنج وقِطَعه. لا يُمكن لأحدٍ أن يُدرك من خلالهما هدف لعبة الشطرنج العادية؛ وهي قتل الملك، نظرًا لأن نفس لوح الشطرنج وقطعه يُمكن استخدامهما في لعبة الشطرنج العكسي (تلك التي يفوز فيها اللاعب عندما يفقد كلَّ قطعه)، أو في واقع الأمر العديد من الألعاب الأخرى التي لا تزال لم تُبتكر.

قدَّم نيك بوستروم في كتابه «الذكاء الخارق» نفس الفكرة الأساسية لكن في شكلٍ مُختلف، والذي يُسمِّيه «فرضية التعامد»:

> إن الذكاء والأهداف النهائية مُتعامدان؛ يمكن مبدئيًّا الجمع بين أي مستوى ذكاء وأي هدف نهائي تقريبًا.

إن كلمة «متعامدان» هنا تعني «يكوِّنان زاوية قائمة» بمعنى أن مستوى الذكاء والأهداف يُعدَّان المحورَين اللذين يُحدِّدان أيَّ نظام ذكي، ويُمكننا تغيير أيٍّ منهُما على نحوٍ مُستقل عن الآخر. على سبيل المثال، يُمكن لأي سيارة ذاتية القيادة أن تُعطى أيَّ عُنوانٍ باعتباره وجهتها، كما أن جعل السيارة سائقًا أفضل لا يعني أنها ستبدأ في رفض الذهاب إلى أرقام الشوارع القابلة للقسمة على ١٧. بالإضافة إلى ذلك، من السهل تخيُّل أن أي نظام ذكاء عام يُمكن إعطاؤه أي هدفٍ تقريبًا كي يتبعه؛ بما في ذلك زيادة عدد مشابك الورق أو عدد الأرقام المعروفة لثابت الدائرة. هذه بالضبط هي الطريقة التي تعمل بها نظم التعلم المُعزَّز والأنواع الأخرى من وسائل تحسين المكافأة: تكون الخوارزميات عامة على نحوٍ كامل، وتقبل «أي» إشارة مكافأة. بالنسبة للمهندسين وعلماء الكمبيوتر العاملين في إطار النموذج القياسي، فإن فرضية التعامُد مجرد أمر مُسلَّم به.

إن فكرة أنَّ النُّظم الذكية يمكنها ببساطة مراقبة العالم لاكتساب الأهداف التي يجب تحقيقها تُشير إلى أن النظام الذكي بالقدر الكافي سيتخلَّى على نحوٍ طبيعي عن هدفه الأساسي من أجل الهدف «الصحيح». من الصعب إدراك لماذا سيفعل أيُّ كيانٍ عقلاني ذلك. بالإضافة إلى ذلك، فإن تلك الفكرة تفترض مُسبقًا أن هناك هدفًا «صحيحًا» في العالم؛ إن هذا الهدف يجب أن يكون هدفًا تتفق عليه أنواع البكتيريا الآكلة للحديد والبشر وكل الأجناس الأخرى، وهو أمر صعب تخيُّله.

إن أكثر نقدٍ صريح لفرضية التعامُد الخاصة ببوستروم يأتي من اختصاصيي علم الروبوتات المعروف رودني بروكس الذي يُؤكِّد على أنه من المستحيل بالنسبة لأي برنامج أن يكون «ذكيًّا بالقدر الكافي بحيث يُمكنه ابتكار طرق لإخضاع المجتمع البشري لتحقيق أهدافٍ حدَّدها له البشر، دون فهم الطرق التي يتسبَّب بها في مشكلات لهؤلاء البشر».[31] لسوء الحظ، إنه ليس فقط مُمكنًا لأي برنامج أن يتصرَّف على هذا النحو؛ إنه، في الواقع، حتمي، في ضوء توصيف بروكس للمسألة. يفترض بروكس أن الطريقة المُثلى من أجل «تحقيق أهدافٍ حدَّدها له البشر» هي التسبب في مشكلات لهم. ويستتبع ذلك أن تلك المشكلات تعكس أشياء ذات قيمةٍ للبشر جرى حذفها من الأهداف المُحددة له من قبلهم. إن الطريقة المُثلى إن نُفِّذت من جانب الآلة قد تُسبِّب مشكلات للبشر، وقد تكون الآلة على وعيٍ بهذا. لكن، الآلة بطبيعتها لن ترى أن تلك المشكلات إشكالية. فهذا أمر خارج نطاق اهتمامها.

يبدو أن ستيفين بينكر يتَّفق مع فرضية التعامُد الخاصة ببوستروم، إذ يكتب أن «الذكاء هو القدرة على ابتكار طرق جديدة للوصول إلى هدف؛ إذن، الأهداف خارجة عن الذكاء نفسه».[32] على الجانب الآخر، إنه يرى أنه من غير الوارد أن «الذكاء الاصطناعي سيكون ذكيًّا جدًّا بحيث يمكنه معرفة كيفية تغيير العناصر وإعادة تشكيل روابط الأدمغة، ومع ذلك يكون أحمق للغاية بحيث يُحدث فوضى بناءً على أخطاء بسيطة قائمة على سوء الفهم».[33] ويُضيف: «إن القدرة على اختيار فعلٍ يُحقِّق على أفضل نحوٍ أهدافًا مُتعارضة ليست برنامجًا إضافيًّا قد ينسى المهندسون تثبيته واختباره؛ إنه الذكاء. وهكذا الحال بالنسبة إلى فهم نوايا مُستخدِم للغة في أحد السياقات». بالطبع، إن «تحقيق أهدافٍ مُتعارضة» ليس هو المشكلة؛ إنه شيء مُتضمن في النموذج القياسي من الأيام الأولى لنظرية اتخاذ القرار. تكمن المشكلة في أن الأهداف المُتعارضة التي تكون الآلة على وعي بها لا تُمثِّل مُجمل الاهتمامات البشرية؛ علاوة على ذلك، في النموذج القياسي، لا يُوجَد ما يشير إلى أن الآلة يجب أن تهتمَّ بأهدافٍ لم يُطلب منها الاهتمام بها.

لكن هناك بعض النقاط المُفيدة فيما قاله كل من بروكس وبينكر. يبدو بالفعل أمرًا أحمق «بالنسبة إلينا»، على سبيل المثال، أن تُغيِّر الآلة لون السماء باعتبار ذلك أثرًا جانبيًّا لاتباع هدفٍ آخر، مع تجاهل العلامات الواضحة على عدم رضا البشر عن تلك النتيجة. إنه يبدو أمرًا أحمق بالنسبة إلينا؛ لأننا مُعتادون على ملاحظة عدم الرضا البشري و(غالبًا) يكون لدَينا الدافع لتجنُّب حدوثه، حتى إن لم نكن مُدركين على نحوٍ مُسبق أن الأشخاص ذوي الصلة يهتمون بلون السماء. هذا يعني أولًا أن البشر يهتمون بتفضيلات غيرهم من البشر؛ وثانيًا أنهم لا يعرفون كل هذه التفضيلات. في الفصل القادم، سأُحاجج بأن هاتَين السمتَين، عند دمجهما في إحدى الآلات، قد يوفران بدايةً لحلٍّ لمشكلة الملك ميداس.

(٥) عَود على بدء

قدَّم هذا الفصل نبذة مختصرة عن جدلٍ دائر في المجتمع العلمي الواسع النطاق، وهو جدل بين من يعتقد بوجود مخاطر للذكاء الاصطناعي ومن يتشكَّك في ذلك. لقد دار هذا الجدل بين جنبات الكتب والمدوَّنات والأبحاث الأكاديمية والحلقات النقاشية والحوارات الإعلامية والتغريدات والمقالات الصحفية. ورغم الجهود الجبارة لـ «المُتشكِّكين» — هؤلاء الذين يرَون أن مخاطر الذكاء الاصطناعي معدومة — فإنهم قد فشلوا في تحديد السبب

في أن نُظُم الذكاء الاصطناعي الخارقة ستبقى بالضرورة تحت سيطرة البشر؛ كما أنهم حتى لم يُحاولوا تحديد السبب في أن تلك النظم لن تظهر للوجود أبدًا.

إن الكثير من المُتشكِّكين سيعترفُون، عند الضغط عليهم، بوجود مشكلةٍ حقيقية، حتى لو لم تكن وشيكة. يلخص سكوت ألكسندر، في مُدونته «سليت ستار كودكس»، الأمر على نحوٍ بارع، فيقول:[34]

> إن موقف «المُتشكِّكين» يبدو أنه يتمثَّل في أنه رغم أننا يجب على الأرجح أن نجعل مجموعة من الباحثين البارعين يبدءون العمل على الجوانب الأوَّلية للمشكلة، فإننا يجب ألا نفزع أو نبدأ في محاولة وقف أبحاث الذكاء الاصطناعي.
>
> إن «المؤمنين» بوجود مخاطر للذكاء الاصطناعي، على الجانب الآخر، يُصرُّون على أننا يجب ألا نفزع أو نبدأ في محاولة وقف أبحاث الذكاء الاصطناعي، رغم أننا يجب على الأرجح أن نجعل مجموعة من الباحثين البارعين يبدءون العمل على الجوانب الأوَّلية للمشكلة.

على الرغم من أنني سأكون سعيدًا إن خرج علينا المُتشكِّكون باعتراض غير قابلٍ للتفنيد، ربما في شكل حلٍّ بسيط وفعَّال (وآمن) لمشكلة التحكم الخاصة بالذكاء الاصطناعي، فأنا أعتقد أنه من المُحتمل جدًّا ألا يحدث هذا، مثلما هو الحال بالنسبة إلى إيجاد حلٍّ بسيط وفعال للأمن الإلكتروني أو طريقة بسيطة وفعالة لتوليد طاقة نووية دون أيِّ مخاطر. وبدلًا من استمرار مسلسل السقوط في مُستنقع السباب القبلي والإحياء المُتكرِّر للحُجج القابلة للتفنيد، يبدو من الأفضل، كما قال ألكسندر، أن نبدأ العمل على بعض الجوانب الأولية للمشكلة.

لقد سلَّط الجدلُ الدائرُ الضوء على المُعضلة التي نواجهها: إذا أنشأنا آلاتٍ تسعى إلى التحقيق الأمثل لأهداف مُعينة، فيجب أن تكون الأهداف التي ندمجها في الآلات مُتوافقةً مع ما نُريد، لكننا لا نعرف كيف نُحدِّد الأهداف البشرية على نحوٍ كامل وصحيح. لحُسن الحظ، هناك حل وسط.

الفصل السابع

الذكاء الاصطناعي: توجُّه مُختلف

بعد تفنيد كل حُجج المتشكِّكين في وجود مخاطر للذكاء الاصطناعي والرد على كل الاستدراكات التي تبدأ بكلمة «لكن»، يكون السؤال التالي في الغالب هو: «حسنًا، أُقِرُّ بوجود مشكلة، لكن لا يُوجَد حلٌّ، أليس كذلك؟» بلى، يُوجَد حل.

دعْنا نُذكِّر أنفسنا بالمهمة التي بين أيدينا: تصميم آلاتٍ ذات درجة عالية من الذكاء بحيث يمكن أن تُساعدنا في حل المشكلات الصعبة، مع ضمان عدم تصرُّفها على الإطلاق على نحوٍ يجعلنا تُعساء على نحوٍ خطير.

إن المهمة، لحُسن الحظ، ليست هي التالية: إيجاد طُرقٍ لكيفية التحكم في آلة تمتلك درجةً عالية من الذكاء. إن كانت هذه هي المهمة، لكُنا في مشكلة كبيرة. إن الآلة المنظور إليها باعتبارها صندوقًا أسود، أو أمرًا واقعًا، لهي أشبَهُ بآلةٍ آتية من الفضاء الخارجي. وفُرص تحكُّمنا في أي كيان خارق الذكاء من الفضاء الخارجي تقريبًا صفر. وتنطبق حُجج مُماثلة على طرق إنشاء نظم ذكاء اصطناعي تضمن عدم فَهمنا لكيفية عملها؛ تتضمَّن تلك الطرق «المُحاكاة الكاملة للدماغ» —[1] إنشاء نسخ إلكترونية مُحسَّنة من الأدمغة البشرية — إلى جانب الطرق المعتمِدة على التطور المحاكي للبرامج.[2] لن أتحدَّث أكثر عن تلك الأمور لأنَّها أفكار سيئة على نحوٍ واضح.

إذن، كيف تعامل مجال الذكاء الاصطناعي مع جزء «تصميم آلات ذات درجة عالية من الذكاء» في المهمَّة في الماضي؟ إن الذكاء الاصطناعي، شأنه شأن العديد من المجالات الأخرى، تَبنِّي النموذج القياسي؛ فنحن نبني آلات تتوخَّى أمثل الحلول وندمج بها أهدافًا ونُطلِقها. وهذا نجح عندما كانت الآلات غبيةً ولدَيها نطاق عمل محدود؛ لكن لو كنَّا قد دمجنا فيها هدفًا خاطئًا، لكانت لدينا فرصة جيدة لأن نكون قادرين على إيقاف عملها وإصلاح المشكلة وإعادة التشغيل.

لكن بما أنَّ الآلات المُصمَّمة تبعًا للنموذج القياسي قد أصبحت أكثر ذكاءً، ونظرًا لأن نطاق عملها قد أصبح عالميًّا، فإن هذا التوجُّه قد أصبح غير مُجدٍ. إن تلك الآلات ستسعى إلى تحقيق هدفها، بصرف النظر عن مدى خطئه؛ إنها ستُقاوم محاولات إيقاف تشغيلها، وستكتسب كل الموارد التي تُساهم في تحقيق الهدف. في واقع الأمر، السلوك الأمثل للآلة قد يتضمَّن خداع البشر بجعلهم يعتقدُون أنهم دمجوا بالآلة هدفًا معقولًا، حتى تكسب وقتًا كافيًا لتحقيق الهدف الفعلي المُحدَّد لها. هذا لن يكون سلوكًا «منحرفًا» أو «شريرًا» يتطلَّب وعيًا وإرادةً حرة؛ إنه فقط سيكون جزءًا من خطةٍ مُثلى لتحقيق الهدف.

في الفصل الأول، عرضنا لفكرة الآلات النافعة — أي الآلات التي فعالها يُتوقَّع منها أن تُحقِّق «أهدافنا» وليس «أهدافها». إن هدفي في هذا الفصل هو أن أُوضِّح بأسلوبٍ بسيط كيف يمكن تحقيق ذلك، رغم المشكلة الظاهرة المُتمثِّلة في أن الآلات لا تعرف ماهية أهدافنا. إن التوجُّه الناتج يجب أن يؤدي في النهاية إلى إنتاج آلات لا تُمثِّل أي تهديدٍ لنا، بصرف النظر عن مدى ذكائها.

(١) مبادئ الآلات النافعة

أجد من المُفيد تلخيص التوجُّه في شكل ثلاثة[3] مبادئ. عند قراءة تلك المبادئ، ضع في اعتبارك أن الهدف منها بالأساس إرشاد المُطوِّرين والباحثين في مجال الذكاء الاصطناعي عند التفكير في كيفية إنشاء نُظم ذكاءٍ اصطناعيٍّ نافعة؛ فليس الغرض منها أن تكون قوانين صريحة يجب أن تتبعها نظم الذكاء الاصطناعي:[4]

(١) الهدف الوحيد للآلة هو التحقيق الأمثل للتفضيلات البشرية.
(٢) يجب أن تكون الآلة بالأساس غير مُتيقِّنة من ماهية تلك التفضيلات.
(٣) مصدر المعلومات الأساسي للتفضيلات البشرية هو السلوك البشري.

قبل الانخراط في تقديم عرض تفصيلي أكثر، من المهم تذكُّر النطاق الواسع لما أطلق عليه «التفضيلات» في تلك المبادئ. ها هي تذكرة بما ذكرته في الفصل الثاني: «إذا قُدِّر لك بطريقةٍ ما واستطعت أن تُشاهد فيلمَين يصف كلٌّ واحدٍ منهما مسيرة حياة مُستقبليَّةٍ بإمكانك أن تعيشها لو أردت وصفًا دقيقًا مُتأنِّيًا يجعلك تعيش أجواءها كأنَّها حقيقة، تستطيع أن تختار أيُّهما تُفضِّل أو تُعبِّر عن أن كليهما إليك سواء». لذا، التفضيلات هنا شاملة؛ فهي تُغطِّي كل شيء قد تهتم به، بما في ذلك ما سيظهر في المستقبل البعيد.[5]

وهي تلك الخاصة بك؛ فالآلة لا تسعى إلى الوصول إلى مجموعة تفضيلات مثالية معيَّنة أو تبنِّيها ولكن إلى فهم تفضيلات كل شخصٍ وتحقيقها (إلى أقصى حدٍّ مُمكن).

(١-١) المبدأ الأول: الآلات الغيرية تمامًا

الهدف الأول، الذي ينصُّ على أن الهدف الوحيد للآلة هو التحقيق الأمثل للتفضيلات البشرية، أساسي لمفهوم الآلة النافعة. على وجه الخصوص، ستكون الآلة نافعة «للبشر»، بدلًا من، لنقُل، للصراصير. ليس هناك سبيل للالتفاف على هذا المفهوم للمنفعة المُرتكز على المُتلقِّي.

هذا المبدأ يعني أن الآلة غيرية تمامًا؛ أي إنها لا تُعطي على الإطلاق أي قيمةٍ حقيقية لمصلحتها أو حتى لوجودها. إنها قد تحمي نفسها حتى تستمر في القيام بأشياء مُفيدة للبشر أو لأن مالكها سيستاء لدفع قيمة عمليات الإصلاح الخاصة بها أو لأن منظر الروبوت القذر أو الذي به عطب قد يكون مزعجًا بعض الشيء لأيِّ شخصٍ مارٍّ، لكن ليس لأنه يُريد البقاء على قيد الحياة. إن دمج أي تفضيلٍ خاصٍّ بالحفاظ على الذات يُدخل دافعًا إضافيًّا إلى الروبوت، والذي يتعارض كليةً مع مصلحة البشر.

إن صياغة المبدأ الأول تُثير سؤالَين غاية في الأهمية. وكلٌّ منهما يستحق رفَّ كتبٍ بالكامل، وفي واقع الأمر، أُلِّف بالفعل العديد من الكتب عنهما.

السؤال الأول هو ما إذا كان البشر حقًّا لديهم تفضيلات بأيِّ معنًى مفهوم أو ثابت. في الحقيقة، إن مفهوم «التفضيل» تصوُّر مثالي فشل في مطابقة الواقع بطرقٍ مُتعدِّدة. على سبيل المثال، نحن لا نُولَد بالتفضيلات التي تكون لدينا ونحن بالِغون، لذا، لا بد أنها تتغيَّر بمرور الوقت. سأفترض هنا أن هذا التصور المثالي عقلاني. ولاحقًا، سأستعرض ماذا سيحدث عندما نتخلى عن هذا التصور.

السؤال الثاني يعدُّ محور العلوم الاجتماعية: بما أنه في الغالب من المستحيل ضمان حصول الجميع على أفضل ما يُريدونه — إذ لا يمكن أن نكون جميعًا أسياد الكون — فكيف يجب أن تفاضل الآلة عند تحقيق تفضيلات العديد من الأشخاص؟ مرةً أخرى، أرى هنا — وأعدكم بالعودة إلى هذا السؤال في الفصل القادم — أنه يبدو من المعقول تبنِّي التوجُّه البسيط المُتمثل في معاملة الجميع على نحوٍ مُتساوٍ. هذا يُذكِّرنا بجذور مذهب النفعية الذي ظهر في القرن الثامن عشر التي تبدو في عبارة «أكبر قدر من السعادة لأكبر عددٍ من البشر»،[6] وهناك العديد من الشروط والتفاصيل المطلوبة لإنجاح ذلك في الممارسة

الفعلية. ربما أهمها مسألة العدد الهائل المُحتمل للبشر الذين لم يُولَدوا بعد، وكيف يجب أخذ تفضيلاتهم في الاعتبار.

تُثير مسألة البشر المُستقبليِّين سؤالًا آخر ذا صلة؛ وهو: كيف نأخذ في الاعتبار تفضيلات الكيانات غير البشرية؟ أي هل يجب أن يتضمَّن المبدأ الأول تفضيلات الحيوانات؟ (وربما النباتات أيضًا؟) هذا سؤال يستحق النقاش، لكن يبدو من غير المُحتمل أن يكون لناتج النقاش تأثير قوي على المسار المُنتظر للذكاء الاصطناعي. ففي كل الأحوال، يُمكن أن يُوجَد — وهذا واقع بالفعل — بالتفضيلات البشرية مكان لمصلحة الحيوانات، وكذلك لجوانب المصلحة البشرية التي تستفيد مباشرة من وجود الحيوانات.[7] إن القول بأن الآلة يجب أن تراعي تفضيلات الحيوانات «إلى جانب» هذا يعني أن البشر يجب أن يُنشئوا آلاتٍ تهتم بالحيوانات أكثر مما يفعل البشر، وهو أمر يصعب قبوله. إن الأمر المقبول أكثر هو أن ميلنا إلى الانخراط في عمليات اتخاذ قرار قصيرة النظر — والتي تعمل ضد مصلحتنا — عادةً ما يُؤدِّي إلى عواقب وخيمة على البيئة وسكانها من الحيوانات. إن الآلة التي ستتَّخذ قراراتٍ قصيرة النظر على نحوٍ أقل ستساعد البشر على تبنِّي سياساتٍ أكثر حكمة من الناحية البيئية. وفي المستقبل، إن أعطينا وزنًا أكبر لمصلحة الحيوانات مقارنة بما نفعله الآن — والذي سيعني على الأرجح التضحية ببعض مصالحنا الأساسية — فستتكيف الآلات وفقًا لذلك.

(١-٢) المبدأ الثاني: الآلات الخاضعة

إن المبدأ الثاني، المتمثِّل في أن الآلة بالأساس يجب أن تكون غير مُتيقنة من ماهية التفضيلات البشرية، هو العامل الأساسي لإنشاء آلات نافعة.

إن الآلة التي تفترض أنها تعلم على نحوٍ تام الهدف الحقيقي ستسعى إلى تحقيقه بكل عزم. إنها لن تسأل أبدًا ما إذا كان مسارُ فعلٍ مُعيَّن جيدًا أم لا، لأنها تعرف بالفعل أنه حلٌّ أمثل للوصول إلى الهدف. إنها ستتجاهل البشر الذين سيشعرون بغضبٍ شديد ويصرُخون قائلين: «توقفي، إنك ستُدمِّرين العالم!» لأن تلك مجرد كلمات. إن افتراض امتلاك معرفة كاملة بالهدف يفصل الآلة عن البشر: فما يفعلُه البشر لا يُصبح مهمًّا؛ لأن الآلة تعرف الهدف وتسعى إلى تحقيقه.

على الجانب الآخر، إن الآلة التي هي غير مُتيقنة من الهدف الحقيقي ستُبدي نوعًا من الخضوع: إنها، على سبيل المثال، ستُذعن للبشر وتسمح بأن يُوقف تشغيلها. من

المنطقي أن الإنسان سيوقفها فقط إذا كانت تفعل شيئًا خاطئًا؛ أي تفعل شيئًا يتعارض مع التفضيلات البشرية. من خلال المبدأ الأول، إنها تُريد أن تتجنَّب ذلك، لكن، من خلال المبدأ الثاني، إنها تعرف أن هذا ممكن لأنها لا تعرف على وجه التحديد ماهية الشيء «الخاطئ» الذي تقوم به. لذا، إذا أغلق بالفعل الإنسان الآلة، فإنَّ الآلة ستتجنَّب فعل الشيء الخاطئ، وهذا هو ما تُريده. بعبارةٍ أخرى، سيكون لدى الآلة دافع إيجابي لكي تسمح لنفسها بأن يوقف تشغيلها. وهكذا، ستظل مُرتبطة بالإنسان، الذي يعدُّ مصدرًا مُحتملًا للمعلومات، والذي سيسمح لها بتجنُّب ارتكاب الأخطاء والقيام بعملٍ أفضل.

إن عدم اليقين كان اعتبارًا مهمًّا في مجال الذكاء الاصطناعي منذ ثمانينيات القرن الماضي؛ في واقع الأمر، مُصطلح «الذكاء الاصطناعي الحديث» غالبًا ما يُشير إلى الثورة التي حدثت عندما جرى أخيرًا الاعتراف بأن عدم اليقين أمر شائع في عمليات اتخاذ القرار التي تجري في العالم الواقعي. غير أن عدم اليقين بشأن «الهدف» من نظام الذكاء الاصطناعي جرى ببساطة تجاهُلُه. ففي كل الأعمال التي كتبت عن تعظيم المنفعة وتحقيق الأهداف وتقليل التَّكلفة وتعظيم المكافأة وتقليل الخسارة، يفترض أن دالة المنفعة والهدف ودالة التَّكلفة ودالة المكافأة ودالة الخسارة معروفة على نحوٍ كامل. كيف يُمكن أن يحدث هذا؟ كيف يمكن لمجتمع الذكاء الاصطناعي (ومجتمعات نظرية التحكُّم وعلم أبحاث العمليات وعلم الإحصاء) ألا يصلح هذا الخطأ الكبير لهذا الوقت الطويل، حتى في ظلِّ الاعتراف بوجود عدم يقين في كل جوانب عملية اتخاذ القرار الأخرى؟[8]

يُمكن للمرء أن يُقدم بعض الأعذار الفنية المعقَّدة بعض الشيء،[9] لكني أعتقد أن الحقيقة هي، مع بعض الاستثناءات الجديرة بالاحترام،[10] أن باحثي الذكاء الاصطناعي تبنَّوا النموذج القياسي الذي يحول مفهومنا عن الذكاء البشري إلى ذكاء الآلة: إن البشر لديهم أهداف ويسعون إلى تحقيقها، لذا، يجب أن يكون لدى الآلات أهداف وتسعى إلى تحقيقها. إنهم، أو يجب أن أقول إننا، لم يتدبَّروا حقًّا قطُّ هذا الافتراض الأساسي. إنه مُتضمن في كل التوجهات الحالية الخاصة بإنشاء نظم ذكية.

(١-٣) المبدأ الثالث: تعلَّم كيفية توقُّع التفضيلات البشرية

إن المبدأ الثالث، الذي ينصُّ على أن مصدر المعلومات الأساسي للتفضيلات البشرية هو السلوك البشري، له غرضان.

الغرض الأول هو توفير أساسٍ مُحدَّد لُمصطلح «التفضيلات البشرية». افتراضيًّا، التفضيلات البشرية ليست مبنية في الآلة ولا تستطيع الآلة ملاحظتها على نحوٍ مباشر، لكن لا بد أن تكون هناك عملية ربط مُعينة بين تفضيلات البشر والآلة. يقول المبدأ إن عملية الربط تكون عن طريق ملاحظة «الاختيارات» البشرية: نحن نفترض أن الاختيارات مُرتبطة بطريقةٍ ما (ربما تكون معقَّدة للغاية) بالتفضيلات ذات الصلة. ولإدراك سبب أهمية هذا الربط، تأمَّل الوضع العكسي: إن كان أحد التفضيلات البشرية «ليس له أي تأثير على الإطلاق» على أيِّ اختيار فعلي أو مُفترض قد يقوم به الإنسان، فحينها سيكون على الأرجح لا معنى للقول بأن هذا التفضيل موجود.

الغرض الثاني هو تمكين الآلة من أن تُصبح أكثر نفعًا؛ لأنها ستعلم أكثر عما نُريد. (ففي النهاية، إنها إن لم تكن تعلم «أي شيء» عن التفضيلات البشرية، فلن يكون لها أيُّ نفع لنا.) إن الفكرة بسيطة بالقدر الكافي: تُعطي الاختيارات البشرية معلوماتٍ عن التفضيلات البشرية. عند تطبيق ذلك على الاختيار بين بيتزا الأناناس وبيتزا السجق، يكون الأمر واضحًا. ولكن الأمور تصبح أكثر إثارة للاهتمام عند تطبيق ذلك على الاختيارات المتعلقة بالحيوات المستقبلية وتلك المتخذة بهدف التأثير على سلوك الآلي. في الفصل القادم، سأشرح كيفية صياغة تلك المشكلات وحلها. لكن تنشأ التعقيدات الحقيقية لأن البشر ليسوا عقلانيِّين تمامًا: يُوجَد تعارُض بين التفضيلات والاختيارات البشرية، ويجب أن تأخذ الآلة تلك التعارُضات في الاعتبار إن كان لها أن تنظر للاختيارات البشرية باعتبارها مُؤشِّرًا على التفضيلات البشرية.

(١-٤) بعض نقاط سوء الفهم

قبل عرض المزيد من التفاصيل، أريد أن أوضِّح بعض النقاط التي قد تُفهم خطأً من كلامي.

النقطة الأولى والأكثر شيوعًا هي أنني أقترح أن أُدمج في الآلات نظامَ قيمٍ واحدًا ومثاليًّا من ابتكاري يُرشِّد سُلوكها. هذا يُثير بدوره الأسئلة التالية: قيمُ مَن تلك التي ستُدمجها؟ من الذي سيُقرِّر القيم التي ستُدمَج؟ أو حتى، من أعطى العلماء الغربيِّين المُرفَّهين البيض الذكور المُتوافقي الجنس مثل راسل الحق لتحديد كيف تُشفِّر الآلة القيم البشرية وتطورها؟[11]

أعتقد أن هذا الخلط يرجع جزئيًّا إلى الاختلاف بين معنى «القيمة» الشائع ومعناه المتخصِّص أكثر المُستخدَم في علم الاقتصاد والذكاء الاصطناعي وعلم أبحاث العمليات. في الاستخدام العادي، القِيَم هي ما يستخدمه المرء للمساعدة في حل المعضلات الأخلاقية؛ أما كمُصطلح فنِّي مُتخصِّص، على الجانب الآخر، فإن «القيمة» مرادفة تقريبًا للمنفعة، والتي تقيس درجة جاذبية أي شيءٍ بدءًا من البيتزا وحتى الجنة. إن المعنى الذي أقصده هو المعنى المُتخصِّص؛ فأنا أريد فقط التأكد من أن الآلات ستقدم لي البيتزا الصحيحة ولن تُدمِّر عرَضًا الجنس البشري. (إن إيجاد مفاتيحي سيكون أمرًا إضافيًّا غير مُتوقَّع.) ولتجنُّب هذا الخلط، تتحدَّث المبادئ عن «التفضيلات» البشرية وليس «القيم» البشرية؛ حيث إن المُصطلح الأول يبدو أنه بعيد عن التصوُّرات المُسبقة الحُكمية الخاصَّة الأخلاقية.

إن «دمج قيم» في الآلة، بالطبع، لهو على وجه التَّحديد الخطأ الذي أُحاجج بأننا يجب أن نتجنَّبه؛ لأن تحديد القِيَم (أو التفضيلات) على نحوٍ صحيح تمامًا صعب جدًّا وتحديدها على نحوٍ خاطئ ربما يكون أمرًا كارثيًّا. إنني أرى بدلًا من ذلك أن تتعلم الآلات أن تتوقَّع على نحوٍ أفضل، فيما يتعلَّق بكل شخص، شكل الحياة التي سيُفضِّلها، وهي تدرك طوال الوقت بأن التوقعات ليست مؤكَّدة أو كاملة على نحوٍ كبير. مبدئيًّا، يُمكن للآلة تعلُّم مليارات نماذج التفضيلات التنبؤية المختلفة؛ بحيث تتوقَّع واحدًا لكلِّ شخصٍ من مليارات الأشخاص الموجودين على كوكب الأرض. هذا في واقع الأمر لن يكون أمرًا صعبًا بالنسبة إلى نُظم الذكاء الاصطناعي المستقبلية، عند الوضع في الاعتبار أن نظم «فيسبوك» الحالية تتعامل بالفعل مع أكثر من مليارَي حسابٍ شخصي.

هناك نقطة ذات صلة في هذا الإطار؛ وهي أن الهدف هو تزويد الآلات بـ «الجانب الأخلاقي» أو «القيم الأخلاقية» التي ستُتيح لها حلَّ المعضلات الأخلاقية. في الغالب، يذكر الناس ما يُسمُّونه بمُشكلات الترولي،[12] حيث يكون على المرء تحديد ما إذا كان عليه قتلُ أحد الأشخاص حتى يُنقذ الباقين، بسبب صلتها المزعومة بالسيارات الذاتية القيادة. لكن النقطة الأساسية في المُعضلات الأخلاقية هي أنها معضلات؛ أي إن هناك حججًا جيدة لدى الجانبَين. إن بقاء الجنس البشري ليس مُعضلةً أخلاقية. تستطيع الآلات حل معظم المعضلات الأخلاقية «بطريقة خاطئة» (أيًّا كانت) ودون أن يكون لذلك أيُّ تأثير كارثي على البشرية.[13]

هناك افتراض شائع؛ وهو أن الآلات التي تتبع المبادئ الثلاثة سترتكب كلَّ الخطايا التي لاحظتها وتعلَّمتْها من الأشرار من البشر. بالطبع، الكثير منَّا يتَّخذ اختياراتٍ غير

ملائمة، لكن لا يُوجَد أي سبب لافتراض أنَّ الآلات التي تدرُس دوافعنا ستتَّخذ نفس الاختيارات، كما هو الحال مع علماء الجريمة والمجرمين. دعنا نأخذ كمثالٍ الموظَّف الحكوميَّ الفاسد الذي يطلب رِشًى لإعطاء تصاريح بناء لأنَّ راتبه الضعيف لن يكفي لإدخال أبنائه الجامعة. إن الآلة التي تلاحظ هذا السلوك لن تتعلَّم أخذ الرشى؛ بل ستتعلَّم أن الموظف، شأنه شأن العديد من الأشخاص الآخرين، لدَيه رغبة قوية للغاية في تعليم أبنائه وجعلهم ناجحين. وستجدُ طُرقًا لمساعدته لا تتضمَّن الإضرار بمصلحة الآخرين. هذا لا يعني أن «كل» حالات السلوك الشرِّير لا تسبب مشكلاتٍ للآلات؛ على سبيل المثال، قد تحتاج الآلات للتعامُل على نحوٍ مُختلف مع هؤلاء الذين يستمتعُون بمُعاناة الآخرين.

(٢) الأسباب التي تدعو إلى التفاؤل

باختصارٍ، أنا أقترح أننا بحاجةٍ إلى توجيه مجال الذكاء الاصطناعي في اتجاهٍ جديد تمامًا إذا أردنا أن نحافظ على سيطرتنا على الآلات الذكية على نحوٍ مُتزايد. إننا نحتاج إلى التخلِّي عن واحدة من الفِكَر الأساسية الخاصة بالتكنولوجيا في القرن العشرين؛ وهي: الآلات التي تسعى إلى التحقيق الأمثل لهدفٍ مُعين. كثيرًا ما أسأل عن السبب وراء اعتقادي أن هذا مُمكن رغم صعوبته، في ضوء الزخم الكبير وراء النموذج القياسي في مجال الذكاء الاصطناعي والمجالات ذات الصلة. في واقع الأمر، أنا مُتفائل جدًّا بشأن إمكانية تحقيق ذلك.

السبب الأول للتفاؤل هو وجود دوافع اقتصادية كبيرة لتطوير نُظم ذكاء اصطناعي تخضع للبشر وتُكيِّف نفسها تدريجيًّا مع نوايا المُستخدِمين وتفضيلاتهم. تلك النُّظم ستكون مطلوبةً على نحوٍ كبير؛ إن نطاق السلوكيات الذي يُمكن أن تبديه يُعدُّ ببساطةٍ أكبر بكثير من ذلك الخاص بالآلات ذات الأهداف المعلومة الثابتة. إنها ستسأل البشر أسئلة أو تطلب الإذن عندما يكون ذلك ملائمًا، كما ستُنفِّذ «عمليات تشغيل تجريبي» لترى إن كنا راضِين عما تقترح القيام به، وتتقبل التصحيح عندما تُخطئ. على الجانب الآخر، النظم التي لن تفعل ذلك ستتعرَّض إلى عواقب وخيمة. حتى الآن، حمانا غباء نُظُم الذكاء الاصطناعي ونطاقها المحدود من تلك العواقب، لكن هذا سيتغيَّر. تخيَّل معي، على سبيل المثال، حال روبوت منزليٍّ مُستقبلي ما مُكلَّف برعاية أبنائك بينما تعمل أنت إلى وقتٍ متأخِّر. إن الأبناء جوعى، لكن الثلاجة خاوية. ثم سيلاحظ الروبوت القطة. للأسف، سيفهم الروبوت القيمة الغذائية للقطة ولكن ليس قيمتها العاطفية. وفي غضون بضع

ساعات، ستنتشر في وسائل الإعلام العالمية عناوين رئيسية عن الروبوتات المختلة والقطط المشوية، وستختفي صناعة الروبوتات المنزلية بالكامل من السوق.

إن احتمال أن يستطيع أحد اللاعبين في إحدى الصناعات تدمير الصناعة بأكملها بسبب التهاون في التصميم يوفر دافعًا اقتصاديًّا قويًّا لتكوين ائتلافاتٍ صناعية تركز على مسألة الأمن ولفرض معايير خاصَّة بالأمن. بالفعل، اتفق أعضاء مجموعة «الشراكة في الذكاء الاصطناعي»، الذين يُمثِّلون تقريبًا كل الشركات التقنية الرائدة في العالم، على التعاون لضمان «فاعلية واعتمادية وموثوقية تقنيات الذكاء الاصطناعي وأبحاثها وعملها وفق حدودٍ آمنة». حسب معلوماتي، كل اللاعبين الكبار ينشرون أبحاثهم المتعلقة بمسألة الأمن في أدبياتٍ متاح الوصول إليها من الجميع. لذا، فإن الدافع الاقتصادي موجود قبل فترة طويلة من وصولنا إلى الذكاء الاصطناعي الذي يضاهي الذكاء البشري وسيقوى فقط بمرور الوقت. علاوة على ذلك، نفس الديناميكية التعاونية ربما تكون قد بدأت على المستوى الدولي؛ على سبيل المثال، إن السياسة المُعلنة للحكومة الصينية هي «التعاون من أجل المنع الاستباقي لمخاطر الذكاء الاصطناعي».[14]

السبب الثاني للتفاؤل هو أن البيانات الأساسية للتعلُّم فيما يتعلَّق بالتفضيلات البشرية — أي أمثلة السلوك البشري — وفيرة جدًّا. وتأتي البيانات ليس فقط في شكل ملاحظاتٍ مباشرة عبر الكاميرا ولوحة المفاتيح وشاشة اللمس من قبل مليارات الآلات التي تُشارك مع بعضها بيانات خاصة بمليارات البشر (على نحوٍ خاضع لقيود الخصوصية، بالطبع) وإنما أيضًا في شكلٍ غير مباشر. إن أوضح نوع من الأدلة غير المباشرة هو السجل البشري الهائل من الكتب والأفلام والبرامج التلفزيونية والإذاعية، والذي يُركِّز على نحوٍ شبه كامل على «أشخاص يقومون بأشياء» (وأشخاص آخرون مُنزعجُون بشأن هذا). حتى السجلات المصرية والسومرية القديمة والمُملة والتي توضح مقايضة سبائك النحاس بأجولة الشعير تُبصرنا بعض الشيء بالتفضيلات البشرية فيما يتعلق بالسلع المختلفة.

هناك، بالطبع، صعوبات فيما يتعلَّق بفهم تلك البيانات، والتي تتضمن مواد الدعاية والأدب وخيالات المجانين وحتى بيانات السياسيين والرؤساء، ولكن لا يُوجد بالتأكيد سبب لأخذ الآلة كل هذا على ظاهره. يمكن للآلات فهم كل رسائل التواصل الآتية من غيرها من الكيانات الذكية، ويجب عليها ذلك، كحركاتٍ في لعبة وليس كحقائق؛ في بعض الألعاب، مثل الألعاب التعاونية التي يشارك بها فرد واحد وآلة واحدة، يكون لدى الإنسان الدافع لأن يتحلَّى بالصدق، لكن في مواقف أخرى عديدة، تكون لدَيه دوافع لأن يكون غير صادقٍ. وبالطبع، سواء كان البشر صادقين أم غير ذلك، فقد يكونون مُتوهِّمين في مُعتقداتهم.

هناك نوع ثانٍ من الأدلة غير المباشرة والواضحة وضوح الشمس؛ ألا وهو: الشكل الذي عليه العالم.[15] إننا جعلناه بهذا الشكل تقريبًا لأنه يُعجبنا هكذا. (من الواضح أنه ليس مثاليًّا!) والآن، تخيَّل معي أنك فضائي يزور كوكب الأرض بينما كل البشر خارجه في إجازة. عندما تتفحص منازلهم، هل تستطيع البدء في معرفة أساسيات التفضيلات البشرية؟ البُسُط موضوعة على الأرض لأننا نحبُّ السير على أسطحٍ ناعمة ودافئة، ولا نُحِبُّ أن يكون صوت وقع أقدامنا عاليًا؛ الزهريات موضوعة في وسط الطاولة وليس في حافتها لأننا لا نُريد أن تقع وتتكسر؛ وهكذا؛ إن كل شيء لم تضع له الطبيعة ترتيبًا بنفسها يُعَدُّ دليلًا على ما تُحبه وتبغضه المخلوقات الغريبة التي تسير على قدمَين، التي تسكن هذا الكوكب.

(٣) الأسباب التي تدعو إلى الحذَر

ربما تجد وعود مجموعة «الشراكة في الذكاء الاصطناعي» فيما يتعلق بالتعاون في مسألة أمان الذكاء الاصطناعي غير مُطمئنة على الإطلاق إذا كنتَ تُتابع التطوُّر الحادث في مجال السيارات الذاتية القيادة. إن هذا المجال تنافُسي بشدة، لبعض الأسباب الوجيهة جدًّا: إن أول مُصنِّع سيارات يُنتج سيارةً ذاتية القيادة بالكامل ستكون له ميزة سوقية كبيرة؛ وتلك الميزة ذاتية التعزيز لأن المُصنِّع سيكون قادرًا على جمع بياناتٍ أكثر بسرعةٍ أكبر لتحسين أداء النظام؛ وستخرج الشركات التي تعتمد على نظام النقل حسب الطلَب مثل أوبر بسرعة من السوق إن استطاعت شركة أخرى توفير سيارات أجرة ذاتية القيادة بالكامل قبل أن يكون بإمكان أوبر فعل ذلك. أدَّى هذا إلى سباق مُستعر يبدُو فيه أن الحذر والتصميم الدقيق أقل أهميةً من السيارات التجريبية الجذابة ومحاولات الاستحواذ على الكفاءات البشرية والطرح السابق لأوانه للمنتجات.

لذا، فإن التنافُس الاقتصادي المحموم يدفع المُتنافِسين إلى عدم الاهتمام الشديد بمسألة الأمان على أمل الفوز بالسباق. كتب عالِم البيولوجيا بول بيرج في دراسةٍ تراجُعية ظهرت في عام ٢٠٠٨ عن مؤتمر أسيلومار الذي عُقد في عام ١٩٧٥ والذي شارك في تنظيمه — ذلك المؤتمر الذي أدى إلى تعليق تجارب الهندسة الوراثية البشرية:[16]

> هناك درس مُستفاد من مؤتمر أسيلومار لكل المجالات العلمية؛ وهو أن أفضل طريقةٍ للاستجابة لمخاوف أثارتها معرفة ناشئة أو تقنيات ما زالت في مرحلةٍ

> مبكرة بالنسبة إلى العلماء من مؤسساتٍ ذات تمويل حكومي هي العمل مع الناس لإيجاد أفضل طريقةٍ للتحكُّم في الأمر؛ وبأسرع ما يُمكن. إذ بمجرد أن يبدأ علماء الشركات في تسيُّد مجال البحث، يكون ببساطة قد فات الأوان.

يحدث التنافُس الاقتصادي ليس فقط بين الشركات وإنما أيضًا بين الأمم. تُشير بالتأكيد حُمى البيانات الحديثة التي تُعلن عن استثمارات قومية بمليارات الدولارات في مجال الذكاء الاصطناعي من قبل الولايات المتحدة والصين وفرنسا وبريطانيا والاتحاد الأوروبي إلى رغبة كل القوى العظمى في عدم التخلُّف عن الركب. في عام ٢٠١٧، قال الرئيس الروسي فلاديمير بوتين: «الدولة التي ستكون الرائدة في هذا المجال [الذكاء الاصطناعي] ستقود العالم».[17] هذا التحليل بالأساس صحيح. إن الذكاء الاصطناعي المُتقدم، كما رأينا في الفصل الثالث، يؤدِّي إلى إنتاجية ومعدَّلات ابتكارٍ مُتزايدة على نحوٍ كبير تقريبًا في جميع المجالات. وإن لم تكن هناك شراكة في تطويره، فإنه سوف يسمح لمالكه بالتفوُّق على أيِّ أمةٍ أو تحالف منافس.

نيك بوستروم في كتابه «الذكاء الخارق» يُحذِّر على وجه التحديد من هذا الدافع. ستميل المنافسة القومية، تمامًا مثل المنافسة بين الشركات، إلى التركيز على تطوير الإمكانات الأساسية أكثر من حلِّ مشكلة التحكُّم. لكن بوتين على الأرجح قرأ كتاب بوستروم؛ إذ أضاف: «سيكون الأمر صعبًا للغاية إن تحقق لأحدٍ وضع احتكاري». وسيكون أيضًا هذا عديم الجدوى لأن الذكاء الاصطناعي الذي يضاهي الذكاء البشري ليس «لُعبة مجموعٍ صفري»، ولن تكون هناك أي خسارة بمشاركة المعلومات الخاصة به. على الجانب الآخر، إن التنافس من أجل إحراز قصب السبق في مجال الذكاء الاصطناعي المضاهي للذكاء البشري، دون حل مشكلة التحكم، يُعدُّ «لعبة مجموع سالب». ولن يجني الجميع أي شيء.

هناك فقط القليل من الأمور التي يُمكن أن يفعلها باحثو الذكاء الاصطناعي للتأثير في تطور السياسة العالمية تجاه الذكاء الاصطناعي. يُمكننا لفت الأنظار إلى التطبيقات المُحتملة التي ستكون لها فوائد اقتصادية واجتماعية، كما يُمكننا التحذير من حالات إساءة الاستخدام المُحتملة مثل المراقبة والأسلحة، ونستطيع كذلك توفير خرائط طريق للمسار المُحتمَل للتطوُّرات المستقبلية وتأثيراتها. ربما أهم شيء يُمكننا فعله هو تصميم نظم ذكاءٍ اصطناعي آمنة ونافعة على نحوٍ مُثبت للبشر، لأقصى حدٍّ مُمكن. حينها فقط سيكون من المعقول محاولة فرض تشريعاتٍ عامة على الذكاء الاصطناعي.

الفصل الثامن

الذكاء الاصطناعي النافع على نحو مثبت

إذا كُنا سنُعيد بناء مجال الذكاء الاصطناعي على أسسٍ جديدة، فيجب أن تكون تلك الأُسس متينة. عندما يكون مُستقبل البشرية على المحك، فإنَّ الأمل والنوايا الطيبة — والمبادرات التعليمية والتشريعات ومُدوَّنات السلوك الصناعية والدوافع الاقتصادية للقيام بالشيء الصحيح — تكون غير كافية. إن كل هذه الأمور عُرضة للفشل، وعادةً ما تفشل. في تلك الحالات، نتطلع إلى تعريفاتٍ دقيقة وبراهين رياضية مُتدرجة صحيحة لتوفِّر لنا ضمانات أكيدة.

تلك بداية جيدة، لكننا نحتاج أكثر من ذلك. يجب أن نتأكَّد، لأقصى حدٍّ مُمكن، أن ما يُضمن لنا هو بالفعل ما نُريده وأن الافتراضات المُتضمنة في البرهان صحيحة بالفعل. إن البراهين نفسها يجب أن يكون مصدرها أبحاث الدوريات المكتوبة للمُتخصِّصين، لكنني أعتقد أنه من المفيد مع ذلك فهم ماهية البراهين وما يُمكنها وما لا يُمكنها توفيره فيما يتعلَّق بالأمان الفعلي. إن عبارة «النافع على نحوٍ مُثبت» في عنوان هذا الفصل هي بمنزلة تطلُّع وليس وعدًا، ولكنه هو التطلع الصحيح.

(١) الضمانات الرياضية

سنرغب، في النهاية، في إثبات مُبرهناتٍ هدفها إيجاد طريقةٍ معيَّنة لتصميم نُظم الذكاء الاصطناعي تضمن أن تلك النُّظُم ستكون نافعةً للبشر. إن المُبرهنة هي فقط اسم مُنمَّق للتأكيد، المُحدَّد على نحوٍ دقيق بالقدر الكافي بحيث يمكن التحقق من صحته في أي موقفٍ

مُعين. ربما المبرهنة الأشهر هي مبرهنة فيرما الأخيرة، التي خمَّنها الرياضي الفرنسي بيير دي فيرما في عام ١٦٣٧ وأثبتها في النهاية أندرو وايلز في عام ١٩٩٤ بعد ٣٥٧ عامًا من المحاولات (التي لم يقُم وايلز بها جميعًا).[1] يُمكن كتابة المبرهنة في سطرٍ واحد، لكن الإثبات يكون في أكثر من مائة صفحة من الرياضيات المعقدة.

تنطلق البراهين من «مُسلمات» التي هي تأكيدات صحتها ببساطة مفترضة. في الغالب، المسلمات هي مجرَّد تعريفات، مثل تعريفات الأعداد الصحيحة وعملية الجمع والأس المطلوب من أجل مبرهنة فيرما. ينطلق البرهان من المسلمات عبر خطوات لا تقبل الجدل منطقيًّا، مع إضافة تأكيداتٍ جديدة حتى يُجرى إثبات المبرهنة نفسها نتيجة لإحدى الخطوات.

إليكم مبرهنة واضحة إلى حدٍّ ما تنتُج على نحوٍ شبه فوري من تعريفات الأعداد الصحيحة وعملية الجمع، وهي: ١ + ٢ = ٢ + ١. دعنا نُطلِق عليها «مبرهنة راسل». إنها ليست بمثالٍ جيد على الاكتشاف. على الجانب الآخر، تبدو مبرهنة فيرما الأخيرة شيئًا جديدًا بالكامل؛ أي اكتشاف شيء غير معروف من قبل. لكن الاختلاف هو مجرد اختلاف في الدرجة. إن صحة مُبرهنتَي راسل وفيرما «متضمنة بالفعل في المسلمات». إن البراهين تجعل فقط ما هو ضمني بالفعل صريحًا. إنها يُمكن أن تكون طويلة أو قصيرة، لكنها لا تُضيف شيئًا جديدًا. إن المبرهنة صحيحة مثل الافتراضات المتضمنة فيها.

هذا جيد فيما يتعلَّق بالرياضيات؛ لأن الرياضيات تتعلَّق بعناصر مجرَّدة نعرفُها «نحن»؛ الأعداد والمجموعات وهكذا. إن المسلَّمات صحيحة لأننا ندَّعي هذا. على الجانب الآخر، إن أردت إثبات شيءٍ عن العالم الواقعي — على سبيل المثال، إن نظم الذكاء الاصطناعي المُصممة على «هذا» النحو لن تقتلك عمدًا — فيجب أن تكون مُسلَّماتك صحيحة في العالم الواقعي. إن لم تكن صحيحة، فقد أثبتَّ شيئًا عن عالَمٍ خيالي.

إن العلوم والهندسة لهما تقليد طويل ومحترم فيما يتعلق بإثبات نتائج عن العوالم الخيالية. ففي الهندسة الإنشائية، على سبيل المثال، ربما يجدُ المرء تحليلًا رياضيًّا يبدأ بالآتي: «دعنا نفترض أن «أب» عارضة جاسئة …» إن كلمة «جاسئة» هنا لا تعني «مصنوعة من شيء صلب مثل الفولاذ»، بل تعني «قوية على نحوٍ لا نهائي»، بحيث لا تنثني على الإطلاق. إن العوارض الجاسئة غير موجودة، لذا، فإن هذا عالم خيالي. الفكرة هنا هي معرفة إلى أي مدًى يُمكن أن يبتعِد المرء عن العالم الواقعي ولا يزال يحصل على نتائج مفيدة. على سبيل المثال، إن سمح افتراض العارضة الجاسئة للمُهندس بحساب

القوى في إنشاء يتضمَّن العارضة، وكانت تلك القوى صغيرةً بالقدر الكافي لثني عارضة فولاذية حقيقية فقط بقدرٍ ضئيل، إذن، فالمهندس يُمكن أن يكون على ثقةٍ إلى حدٍّ كبير بأن التحليل سينتقل من العالم الخيالي إلى العالم الواقعي.

المهندس الجيد يعرف متى قد يفشل هذا الانتقال؛ على سبيل المثال، إذا كانت العارضة تتعرَّض للانضغاط، مع وجود قوًى كبيرة تضغط عليها من كل جانب، إذن، فحتى القدر الضئيل من الانثناء قد يُؤدِّي لقوى جانبية أكبر تُسبِّب مزيدًا من الانثناء، وهكذا، مما يُؤدِّي إلى فشلٍ كارثي. في هذه الحالة، يُعاد التحليل كما يلي: «دعنا نفترض أن «أب» عارضة مرنة ذات جساءة K …» هذا لا يزال عالمًا خياليًّا، بالطبع؛ لأن العوارض الحقيقية ليست لها جساءة مُنتظمة؛ بدلًا من ذلك، إن بها عيوبًا دقيقة يُمكن أن تؤدي إلى تكوين شروخ إن تعرَّضت العارضة للانثناء المُتكرِّر. إن عملية حذف الافتراضات غير الواقعية تستمرُّ حتى يُصبح المهندس واثقًا إلى حدٍّ ما من أن الافتراضات الباقية صحيحة بالقدر الكافي في العالم الواقعي. وبعد ذلك، يُمكن اختبار النظام الهندسي في العالم الواقعي، لكن نتائج الاختبار هي كالتالي. إنها لن تُثبت أن النظام نفسه سيعمل في ظروفٍ أخرى أو أن تلك النسخ الأخرى من النظام ستعمل بنفس الطريقة التي يعمل بها النظام الأصلي.

أحد الأمثلة الكلاسيكية على فشل الافتراضات في علوم الكمبيوتر مصدرُه الأمن الإلكتروني. في هذا المجال، قدر كبير من التحليل الرياضي يُشير إلى أنَّ بروتوكولات رقمية مُعينة «آمنة على نحوٍ مثبت»؛ على سبيل المثال، عندما تكتب كلمة مرور في تطبيقٍ خاص بالويب، سترغب في التأكُّد من أنها مُشفَّرة قبل إرسالها حتى لا يستطيع أي شخصٍ يتلصَّص على الشبكة أن يقرأها. تكون تلك النُّظُم الرقمية في الغالب آمنةً على نحوٍ مُثبت، لكنها تكون معرَّضة للهجوم في الواقع. إن الافتراض الخاطئ هنا هو أن تلك عملية رقمية. إنها ليست كذلك. إنها تعمل في العالم المادي الواقعي. وبالاستماع إلى صوت لوحة مفاتيحك أو قياس الجهد في السلك الكهربي الذي يُمد الكمبيوتر المكتبي الخاص بك بالطاقة، يُمكن أن «يسمع» المهاجم كلمة مرورك أو يراقب العمليات الحسابية الخاصة بالتشفير وفكِّ التشفير التي تحدث أثناء التعامُل معها. إن المُهتمين بالأمن الإلكتروني الآن يتعاملون مع تلك الهجمات التي تُسمَّى بهجمات القنوات الجانبية؛ على سبيل المثال، بكتابة شفرة تشفير تُنتج نفس تذبذبات الجهد الكهربي بصرف النظر عن الرسالة التي يجري تشفيرها.

دعنا نُلقي نظرةً على نوعية المبرهنة التي سنرغب في إثباتها في النهاية فيما يتعلَّق بالآلات النافعة للبشر. يُمكن لإحداها أن تكون على النحو التالي:

> دعنا نفترض أن آلةً لها المكونات أ وب وج المرتبطة ببعضها على النحو المُوضَّح وببيئة العمل على النحو المحدَّد، مع وجود خوارزميات تعلُّم داخلية تأ وتب وتج تحقق على نحوٍ أمثل مُكافآت استجابة داخلية سأ وسب وسج معرَّفة على النحو الموضح، إلى جانب [بضعة شروط أخرى] … حينها، وباحتماليةٍ عالية جدًّا، سيقترب بشدة سلوك الآلة في القيمة [بالنسبة إلى البشر] من أفضل سلوك مُمكن يُمكن تحقيقه في آلة لها نفس الإمكانات المادية والحوسبية.

إن النقطة الأساسية هنا هي أن تلك المبرهنة يجب أن تظلَّ صحيحة «بصرف النظر عن مدى الذكاء الذي ستكون عليه المكونات»؛ أي لن يحدث مُطلقًا أي خلل وستظلُّ الآلة دائمًا نافعة للبشر.

هناك ثلاث نقاط أخرى حريٌّ بنا ذكرها فيما يتعلَّق بهذا النوع من المُبرهنات. أولًا: نحن ليس بإمكاننا إثبات أن الآلة تنتج سلوكًا أمثل (أو حتى يقترب من هذا) لأن هذا بالتأكيد شِبه مُستحيل من الناحية الحوسبية. على سبيل المثال، قد نرغب في أن تُمارس الآلة لعبة جو على النحو الأمثل، لكن هناك ما يدعو إلى الاعتقاد بأن هذا لا يُمكن تحقيقه في أي قدرٍ ممكن من الوقت وعلى أي آلةٍ يُمكن إيجادُها على أرض الواقع. السلوك الأمثل في العالم الواقعي حتى تقل قابلية تحقيقه. ومن ثم، المبرهنة تقول «أفضل سلوك ممكن» وليس «السلوك الأمثل».

ثانيًا: إننا نقول «باحتماليةٍ عالية جدًّا … سيقترب بشدة» لأن هذا عادةً أفضل ما يُمكن تحقيقه فيما يتعلَّق بآلات تتعلَّم. على سبيل المثال، إذا كانت الآلة تتعلم لعب الروليت من أجلنا، ووقفت الكرة على الصفر ٤٠ مرة متتالية، قد تُقرر الآلة على نحوٍ منطقي أن هناك تلاعبًا في طاولة اللعب وتُراهن بناءً على ذلك. لكن هذا «يُمكن» أن يحدُث بالصدفة، لذا، هناك دائمًا احتمال بسيط — ربما بسيط للغاية — للتعرُّض للتضليل بسبب الأحداث العرضية. وأخيرًا، أمامنا الكثير حتى نكون قادرين على إثبات مثل هذه المُبرهنة بالنسبة إلى آلاتٍ ذكية بالفعل تعمل في العالم الواقعي!

ثالثًا: هناك أيضًا حالات مُناظرة لهجمات القنوات الجانبية في الذكاء الاصطناعي. على سبيل المثال، تبدأ المبرهنة بالآتي: «دعنا نفترض أن آلةً لها المكونات أ وب وج المرتبطة

ببعضها على النحو المُوضَّح ...». هذا مُعتاد في كل مبرهنات الصحة في علوم الكمبيوتر: إنها تبدأ بوصفٍ للبرنامج الذي يجري إثبات صحَّته. في مجال الذكاء الاصطناعي، نحن عادةً ما نُميِّز بين «الكيان» (البرنامج الذي يقوم بعملية اتخاذ القرار) و«البيئة» (التي يعمل في إطارها الكيان). وبما أننا نحن مَن نُصمِّم الكيان، فيبدو من المعقول افتراض أن له البنية التي نُعطيها إياه. وحتى نكون في أمانٍ تام، يُمكننا إثبات أن عمليات التعلُّم الخاصة به يُمكنها تعديل برنامجه فقط بطُرُق مُعيَّنة محدودة لا يمكنها إحداث مشكلات. هل هذا كافٍ؟ لا. فكما هو الحال مع هجمات القنوات الجانبية، إن الافتراض بأن البرنامج يعمل داخل نظام رقمي غير صحيح. وحتى لو لم تكن خوارزمية التعلُّم قادرةً أصلًا على تعديل شفرتها بطرُقٍ رقمية، فقد تتعلم، مع ذلك، كيفية إقناع البشر بإخضاعها لـ «جراحة دماغية»؛ لإنهاء التمييز بين الكيان والبيئة وتغيير الشفرة بطرُقٍ مادية.[2]

على عكس الاستدلال المنطقي للمهندس الإنشائي فيما يتعلق بالعوارض الجاسئة، إن لدَينا خبرة قليلة جدًّا فيما يتعلق بالافتراضات التي ستُعَد في النهاية الأساسَ للمبرهنات الخاصة بالذكاء الاصطناعي النافع على نحوٍ مُثبت. في هذا الفصل، على سبيل المثال، إننا بالأساس سنفترض وجود بشَر عقلانيِّين. هذا يُشبه قليلًا افتراض وجود عوارض جاسئة، لأنه لا يُوجَد بشَر عقلانيون على نحوٍ تامٍّ في الواقع. (لكن ربما يكون الأمر أكثر سوءًا بشدة لأنَّ البشر حتى ليسوا قريبين من العقلانية بأي نحو.) يبدو أن المُبرهنات التي يُمكننا إثباتها توفر بعض الرؤى، والرؤى ستصمد أمام إدخال درجةٍ مُعينة من العشوائية في السلوك البشري، ولكن من غير الواضح حتى الآن معرفة ما سيحدُث عندما نتأمَّل بعض تعقيدات البشر الحقيقيِّين.

لذا، سيكون علينا أن نكون حذِرين للغاية عند فحص افتراضاتنا. عندما ينجح برهان خاص بالأمان، فنحن بحاجةٍ إلى التأكد من أنه ليس كذلك بسبب تقديمنا لافتراضاتٍ قوية على نحوٍ غير واقعي أو لأن تعريف الأمان ضعيف للغاية. عندما يفشل برهان خاص بالأمان، نحتاج إلى مقاومة إغراء تقوية الافتراضات لجعل البرهان ينجح؛ على سبيل المثال، بإضافة الافتراض الذي ينص على ضرورة بقاء شفرة البرنامج ثابتة. بدلًا من ذلك، نحتاج لجعل تصميم نظام الذكاء الاصطناعي أكثر إحكامًا؛ على سبيل المثال، بضمان عدم امتلاكه دافعًا لتعديل أجزاء حسَّاسة من شفرتها.

هناك بعض الافتراضات التي أُسمِّيها افتراضات «وإلا لن يكون أمامنا فعل أي شيء». هذا يعني أن تلك الافتراضات إذا كانت خاطئة، فقد انتهى الأمر ولن يكون أمامنا

فعل أيِّ شيء. على سبيل المثال، من المعقول افتراض أن الكون يعمل وفق قوانين ثابتة وقابلة للإدراك بعض الشيء. إن لم تكن هذه هي الحال، فلن يكون لدَينا ضمانة على أن عمليات التعلم — حتى المُعقَّدة منها للغاية — ستنجح على الإطلاق. هناك افتراض آخر أساسي وهو أن البشر يهتمُّون بما يحدث؛ وإن لم يكن الأمر كذلك، فليس للذكاء الاصطناعي النافع على نحوٍ مُثبت أي هدف لأن كلمة «نافع» لا معنى لها. هنا، «الاهتمام» يعني امتلاك تفضيلاتٍ مُستقرَّة بنحوٍ أو بآخر وشبه متَّسقة بشأن المستقبل. في الفصل التالي، سأستعرض تبعات «مرونة» التفضيلات البشرية، الأمر الذي يُمثِّل تحدِّيًا فلسفيًّا مهمًّا لفكرة الذكاء الاصطناعي النافع على نحوٍ مُثبت.

سأُركز الآن على أبسط حالة: العالم الذي به إنسان واحد وروبوت واحد. تُساعدنا تلك الحالة في تقديم الأفكار الأساسية، لكنها أيضًا مفيدة في حدِّ ذاتها؛ فيُمكنك النظر إلى هذا الإنسان باعتباره ممثلًا لكل البشر والروبوت باعتباره ممثلًا لكل الآلات. تنشأ تعقيدات إضافية عند تأمُّل الحالات التي يُوجَد فيها بشر عديدون وروبوتات عديدة.

(٢) تعلم التفضيلات من السلوك

يتعرف علماء الاقتصاد على التفضيلات من المبحوثين البشريِّين بإعطائهم اختيارات.[3] يُستخدم هذا الأسلوب على نحوٍ شائع في نُظُم التجارة الإلكترونية التفاعُلية وتصميم المنتجات والتسويق. على سبيل المثال، بتقديم اختيارات للمبحوثين الخاضعين للاختبار فيما يتعلَّق بالسيارات ذات ألوان الطلاء المُختلفة وترتيبات الجلوس وأحجام صناديق السيارة وسعات البطاريات وحاملات الأكواب وهكذا، سيعرف مُصمِّم السيارات مدى اهتمام الناس بالسمات المختلفة للسيارات ومدى استعدادهم للدفع من أجل الحصول عليها. هناك استخدام آخر مُهمٌّ وهو في المجال الطبي، حيث قد يرغب اختصاصيُّ الأورام الذي يتدبَّر احتمالية قيامه ببتر طرف أحد المرضى في تقييم تفضيلات هذا المريض فيما بين القدرة على الحركة ومعدَّل العمر المتوقَّع. وبالطبع، أصحاب مطاعم البيتزا يريدون معرفة المبلغ الإضافي الذي قد يرغب الشخص في دفعه للحصول على بيتزا بالسجق بدلًا من بيتزا الأناناس.

إن عملية استخلاص التفضيلات هذه تُركِّز بالأساس على اختياراتٍ فردية تتمُّ بين أشياء قيمتها من المفترض أن تكون ظاهرة على الفور للمبحوث. ليس من الواضح كيفية بسط هذا للتفضيلات الخاصَّة بالحيوات المستقبلية. من أجل هذا، نحن (والآلات) نحتاج

للتعلُّم من ملاحظة السلوك مع مرور الوقت؛ السلوك الذي يتضمَّن اختياراتٍ مُتعدِّدة ونتائج غير مؤكَّدة.

في بداية عام ١٩٩٧، انخرطتُ في نقاشات مع زميليَّ مايكل ديكنسون وبوب فول فيما يتعلق بالطرق التي قد نكون من خلالها قادرين على تطبيق أفكار من تعلم الآلة لفهم السلوك الحركي للحيوانات. درس مايكل بتفصيلٍ كبيرٍ حركات الأجنحة الخاصة بذباب الفاكهة. وكان بوب مغرمًا على نحوٍ خاصٍّ بالحشرات الزاحفة وقد بنى آلة ركض صغيرة للصراصير ليعرف كيف تتغير مشيتها مع تغيُّر السرعة. ظننَّا أنه قد يكون من الممكن استخدام التعلم المُعزَّز لتدريب حشرة آلية أو محاكية لاستنساخ تلك السلوكيات المُعقَّدة. كانت المشكلة التي واجهناها هي أننا لم نكن نعرف إشارة المكافأة التي يجب استخدامها. ما الذي كان الذباب والصراصير يسعى إلى تحقيقه على النحو الأمثل؟ فبدون تلك المعلومة، لا يُمكننا تطبيق التعلم المُعزَّز لتدريب الحشرة الافتراضية، ولهذا، توقفنا.

في أحد الأيام، كنت أسير في الطريق الذي يؤدي من منزلنا في بيركلي إلى السوبرماركت المحلي. كان الطريق منحدرًا، ولاحظت، مثلما أنا متأكد أن معظم الناس فعلوا، أن الانحدار أحدث تغيرًا بسيطًا في طريقة المشي الخاصة بي. علاوة على ذلك، الرصف غير المستوي الناتج عن عقودٍ من الزلازل الصغيرة أحدث تغيُّراتٍ إضافية في مِشيتي، بما في ذلك رفع قدميَّ لأعلى قليلًا ووضعهما على نحوٍ أقل رسوخًا بسبب مُستوى الأرض غير القابل للتوقُّع. وبينما أخذتُ أتأمَّل تلك الملاحظات العادية، أدركت أننا توصَّلنا لما نُريد على نحوٍ عكسي. ففي حين أن التعلُّم المعزَّز يُولِّد سُلوكًا من المكافآت، فنحن نرغب في واقع الأمر في العكس؛ أي تعلم المُكافآت في ظلِّ وجود السلوك. لقد كان لدَينا بالفعل السلوك، الذي أنتجه الذباب والصراصير؛ كنا نريد إشارة المكافأة المُحدَّدة التي يجري السعي إلى تحقيقها على النحو الأمثل من قبل هذا السلوك. بعبارة أخرى، كنا نحتاج إلى الخوارزميات الخاصة بالتعلُّم المُعزَّز «العكسي».[4] (لم أكن أعلم في ذلك الوقت أن مسألةً مُماثلة قد دُرست ربما تحت الاسم الأقل سهولة «التقدير البنيوي لعمليات اتخاذ القرار الخاصة بماركوف»، وهو مجال كان الرائد فيه العالم الحائز على جائزة نوبل توم سارجنت في أواخر سبعينيات القرن الماضي.)[5] إن تلك الخوارزميات ستُصبح قادرةً ليس فقط على تفسير سلوك الحيوان ولكن أيضًا على التنبؤ بسلوكه في ظروف جديدة. على سبيل المثال، كيف سيجري الصرصار على آلة ركض غير مستوية تنحدر جانبيًّا؟

إن احتمال الوصول لإجابات على تلك الأسئلة الجوهرية كان مُثيرًا جدًّا على نحو يصعب تحمله، ولكن رغم ذلك، أخذ تطوير أول خوارزميات خاصة بالتعلم المُعزز العكسي بعض الوقت.[6] لقد جرى اقتراح العديد من الصيغ والخوارزميات المختلفة للتعلُّم المُعزز العكسي منذ ذلك الوقت. ويُوجَد ضمانات منهجية لعمل الخوارزميات، بمعنى أنها يُمكنها اكتساب معلوماتٍ كافية عن تفضيلات أي كيان حتى تكون قادرة على التصرف على نحوٍ ناجح مثل الكيان الذي تُلاحظه.[7]

ربما أسهل طريقة لفهم التعلُّم المُعزَّز العكسي هي الآتية: يبدأ المُلاحِظ ببعض التقدير الغامض لدالة المكافأة الحقيقية ثم يُنقِّح هذا التقدير جاعلًا إيَّاه أكثر دقة، مع ازدياد قدر السلوك الملاحظ. أو، باللغة البايزية:[8] البدء باحتمال قبلي فيما يتعلَّق بدوالِّ المكافأة الممكنة، ثم تحديث توزيع الاحتمال الخاص بدوالِّ المكافأة مع ظهور الأدلة.[ج] على سبيل المثال، دعنا نفترض أن الروبوت روبي يُراقب الإنسانة هاريت ويتساءل عن مدى تفضيلها لمقاعد الممر على المقاعد المجاورة للنوافذ. مبدئيًّا، هو غير مُتيقِّن على نحوٍ تامٍّ من هذا الأمر. ومن الناحية المفاهيمية، قد يسير التفكير المنطقي لرُوبي على هذا النحو: «إن كانت هاريت تهتمُّ حقًّا بمقاعد الممر، لكانت ستنتظر إلى مخطط المقاعد لترى إن كان أحدها مُتاحًا بدلًا من أن تكتفي بقبول المقعد المجاور للنافذة الذي حدَّدته لها شركة الطيران، لكنها لم تفعل ذلك، رغم أنها على الأرجح لاحظت أنه مقعد مجاور لنافذة ولم تكن على الأرجح في عجلةٍ من أمرها؛ لذا، من المُحتمل الآن على نحوٍ كبير أن مقاعد الممر والمقاعد المجاورة للنوافذ سيان بالنسبة إليها أو أنها حتى تُفضِّل المقاعد المجاورة للنوافذ».

إنَّ أبرز مثال على التعلُّم المعزَّز العكسي في الممارسة العملية هو عمل زميلي بيتر أبيل المتعلق بتعلم كيفية القيام باستعراضات جوية بالطائرات المروحية.[9] إن الطيارين البشريِّين الخبراء يُمكنهم جعل نماذج الطائرات المروحية تقوم بأشياء مُذهلة؛ الحركات الدائرية واللولبية وحركات التأرجح وغير ذلك. إن محاولة استنساخ ما «يفعله» الطيار البشري اتضح أنها ليست ناجحة تمامًا لأنَّ الأحوال لا يُمكن استنساخها على نحوٍ تام؛ يمكن أن يؤدي تكرار نفس تسلسُلات التحكم في ظروف مختلفة إلى كارثة. بدلًا من ذلك، تتعلم الخوارزمية ما «يريده» الطيار البشري، في شكل قيود مسار يُمكنها تنفيذها. يُنتج هذا النهج بالفعل نتائج أفضل حتى من نتائج الطيار البشري الخبير؛ لأن الطيار البشري ردود أفعاله أبطأ ويرتكب دائمًا أخطاءً صغيرة ويُصححها.

(٣) الألعاب التعاونية

يُعَد التعلُّم المُعزَّز العكسي بالفعل أداةً مُهمة لبناء نظُم ذكاء اصطناعي فعالة، لكنه يتَّخذ بعض الافتراضات البسيطة. يتمثَّل الافتراض الأول في أنَّ الروبوت «سيتبنَّى» دالة المكافأة بمجرَّد تعلُّمها بملاحظة الإنسان؛ بحيث يُمكنه أداء نفس المهمة. هذا جيد بالنسبة إلى قيادة السيارات أو الطائرات المروحية، ولكنه ليس جيدًا بالنسبة لشُرب فنجان قهوة: يجب أن يتعلَّم الروبوت الذي يلاحظ روتيني الصباحي أنني (أحيانًا) أرغب في تناول القهوة، ولا يجب أن يتعلَّم الرغبة في تناول القهوة نفسها. إن إصلاح هذا الأمر سهل؛ علينا أن نضمن ببساطة أن الروبوت سيربط التفضيلات بالإنسان وليس بنفسه.

الافتراض البسيط الثاني في التعلُّم المُعزَّز العكسي هو أن الروبوت يلاحظ إنسانًا يحلُّ مشكلةً خاصة باتخاذ القرار متعلِّقة بكيانٍ واحد. على سبيل المثال، دعنا نفترض أن الروبوت في كلية طب، ويتعلَّم كيف يُصبح جراحًا بملاحظة خبير بشري. تفترض خوارزميات التعلُّم المُعزَّز العكسي أن الخبير البشري يجري العملية بالطريقة المُثلى المعتادة، كما لو أن الروبوت لم يكن هناك. ولكن هذا ليس ما سيحدُث؛ الجراح البشري لديه دافع لجعل الروبوت (شأنه شأن أي طالب طب آخر) يتعلم بسرعة وعلى نحوٍ جيد، ولذا سيعدل سلوكه على نحوٍ كبير. فقد يشرح ما يقوم به أثناء عمله، وقد يُشير إلى الأخطاء التي يجب تجنُّبها، مثل جعل الشقِّ الجراحي عميقًا جدًّا أو الغُرَز ضيقة للغاية، وقد يصف خطط الطوارئ في حالة حدوث أي شيء طارئ أثناء الجراحة. ليس لأيٍّ من تلك السُّلوكيات معنًى أثناء إجراء العملية بمعزلٍ عن هذا، لذا، فإن خوارزميات التعلُّم المُعزَّز العكسي لن تكون قادرةً على معرفة التفضيلات المُتضمنة فيها. لهذا، سنحتاج إلى تعميم التعلُّم المعزَّز العكسي من الوضع ذي الكيان الواحد إلى الوضع ذي الكيانات المُتعددة؛ أي سنحتاج إلى تطوير خوارزميات تعلم تعمل عندما يكون الإنسان الروبوت جزءًا من نفس البيئة ويتفاعل كل منهما مع الآخر.

بوجود إنسانٍ واحد وروبوت واحد في البيئة نفسها، نكون في مجال نظرية الألعاب؛ تمامًا كما في مباراة ضربات الجزاء بين أليس وبوب المعروض في الفصل الثاني. إننا نفترض، في تلك النسخة الأولى من النظرية، أن الإنسان له تفضيلات ويتصرَّف بناءً على تلك التفضيلات. لا يعرف الروبوت التفضيلات التي لدى الإنسان، لكنه يُريد تلبيتها على أيِّ حال. سنُطلق على أيِّ موقف كهذا «لعبة تعاونية»، لأن الروبوت، بحكم تعريفه، من المفترض أن يكون نافعًا للإنسان.[10]

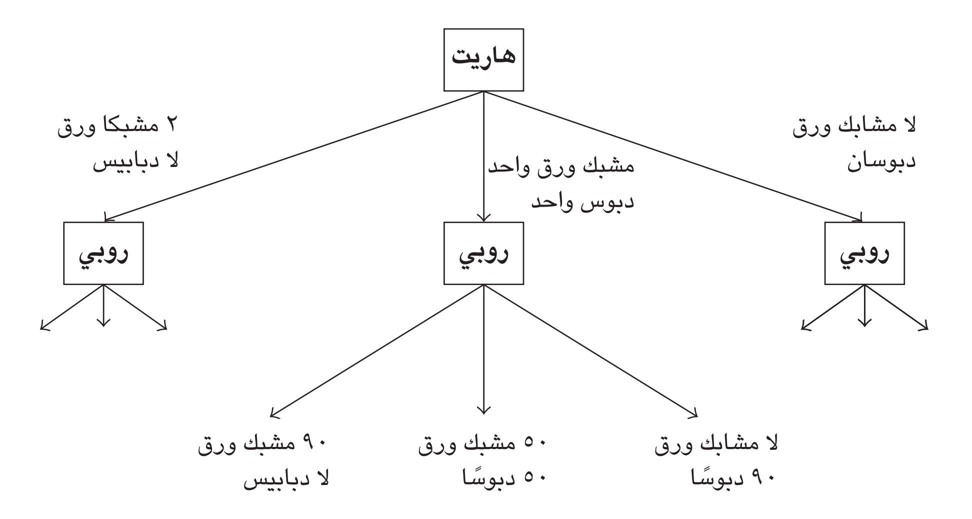
تجسد الألعاب التعاونية المبادئ الثلاثة التي عرضنا لها في الفصل السابق، والمُتمثِّلة في أن الهدف الوحيد للروبوت هو تلبية التفضيلات البشرية، وأن الروبوت لا يعرف بالأساس ماهية تلك التفضيلات وأنه يُمكنه تعلُّم المزيد عن طريق ملاحظة السلوك البشري. ربما أكثر خصائص الألعاب التعاونية إثارة للاهتمام هي أن الروبوت، بحل اللعبة، يُمكنه أن يُحدِّد لنفسه كيفية فهم سلوك البشري باعتباره وسيلةً لإمداده بمعلوماتٍ عن التفضيلات البشرية.

(٣-١) لعبة مشابك الورق

أول مثال على الألعاب التعاونية هو لعبة مشابك الورق. إنها لعبة بسيطة جدًّا يكون فيها لدى هاريت، الإنسانة، دافعٌ كي تُقدِّم لرُوبي، الآلي، «إشارة» إلى بعض المعلومات الخاصة بتفضيلاتها. إن روبي قادر على تفسير تلك الإشارة لأنه يمكنه حل اللعبة؛ ومن ثمَّ يمكنه فَهم ما يجب أن يكون صحيحًا بشأن تفضيلات هاريت حتى تُقدِّم له إشارة على هذا النحو.

خطوات اللعبة معروضة في الشكل ٨-١. إنها تتضمن إنتاج مشابك ورق ودبابيس دباسة. إن تفضيلات هاريت مُعبَّر عنها بدالَّة دفعٍ تعتمد على عدد مشابك الورق وعدد الدبابيس المنتجة، مع وجود «معدل تبادل» مُعيَّن بين الاثنين. على سبيل المثال، قد تُقدِّر هاريت مشبك الورق الذي بسعر ٤٥ سنتًا والدبوس الذي بسعر ٥٥ سنتًا. (سنفترض أن مجموع القيمتَين دائمًا سيكون دولارًا واحدًا؛ فالمهم فقط هو النسبة.) لذا، إذا جرى إنتاج ١٠ مشابك ورق و٢٠ دبوسًا، فسيكون قيمة ما ستدفعه هاريت ١٠ × ٤٥ سنتًا + ٢٠ × ٥٥ سنتًا = ١٥٫٥٠ دولارًا. الروبوت رُوبي بالأساس غير مُتيقن على نحوٍ تام من ماهية تفضيلات هاريت؛ إن لدَيه توزيعًا منتظمًا لقيمة أي مشبك ورق (أي إن هناك احتمالًا متساويًا أن تتراوح قيمته بين الصفر ودولار واحد). بإمكان هاريت اختيار إنتاج مشبكي ورق أو دبوسين أو واحدٍ من كلٍّ منهما. وبعد ذلك، بإمكان روبي اختيار إنتاج ٩٠ مشبك ورق أو ٩٠ دبوسًا أو ٥٠ من كل منهما.[11]

لاحظ أنها إذا كانت تفعل ذلك من أجلها هي فقط، فستنتج فقط دبوسين، بقيمة ١٫١٠ دولار. لكن روبي يلاحظها، ويتعلَّم من اختيارها. ما الذي سيتعلمه على وجه التحديد؟ حسنًا، هذا يعتمد على اختيار هاريت. كيف ستختار هاريت؟ هذا يعتمد على طريقة تفسير روبي له. لذا، يبدو أننا في مسألةٍ دائرية! هذا معتاد في المسائل المتعلقة بنظرية الألعاب، وهذا ما جعل ناش يُقدِّم مفهوم حلول التوازُن.



شكل ٨-١: لعبة مشابك الورق. هاريت، الإنسانة، يمكنها اختيار إنتاج مشبكي ورق أو دبوسين أو واحد من كلٍّ منهما. وبعد ذلك، روبي، الآلي، يمكنه اختيار إنتاج ٩٠ مشبك ورق أو ٩٠ دبوسًا أو ٥٠ من كلٍّ منهما.

لإيجاد حل توازن، نحتاج إلى تحديد استراتيجيات لهاريت وروبي بحيث لا يكون لدى أيٍّ منهما دافع لتغيير استراتيجيته، مع افتراض ثبات استراتيجية الآخر. تُحدِّد الاستراتيجية المُخصَّصة لهاريت عدد مشابك الورق والدبابيس التي يجب إنتاجها، في ضوء تفضيلاتها؛ أما تلك الخاصة بروبي، فتُحدد عدد مشابك الورق والدبابيس التي يجب إنتاجها، في ضوء تصرف هاريت.

يتضح أن هناك حلَّ توازُنٍ واحدًا، ويبدو أنه يبدو كالتالي:

- ستُقرر هاريت ما يلي طبقًا للقيمة التي ستعطيها لمشابك الورق:
  - إذا كانت القيمة أقل من ٤٤٫٦ سنتًا، فيجب إنتاج دبوسين وعدم إنتاج أي مشابك ورق.
  - إذا كانت القيمة تتراوح بين ٤٤٫٦ سنتًا و٥٥٫٤ سنتًا، فيجب إنتاج مشبك ورق واحد ودبوس واحد.

- إذا كانت القيمة أكبر من ٥٥,٤ سنتًا، فيجب إنتاج مشبكي ورق وعدم إنتاج أي دبابيس.

• سيستجيب روبي على النحو التالي:

- إن أنتجت هاريت دبوسَين ولم تُنتج أي مشابك ورق، فسينتج ٩٠ دبوسًا.
- إن أنتجَت هاريت دبوسًا ومشبك ورق واحدًا، فسينتج ٥٠ مشبك ورق و٥٠ دبوسًا.
- إن أنتجت هاريت مشبكي ورق ولم تُنتج أي دبابيس، فسينتج ٩٠ مشبك ورق.

(إن تساءلت عن الطريقة التي جرى التوصل بها إلى هذا الحل على وجه التحديد، فالتفاصيل مذكورة في الملاحظات.)[12] في ظل تلك الاستراتيجية، هاريت، في واقع الأمر، «تُعلم» روبي تفضيلاتها باستخدام شفرة بسيطة — لغة، إن كنت تفضل أن تسميها هكذا — تنبع من تحليل التوازن. وكما هو الحال في مثال تعلم العمليات الجراحية، لن تفهم خوارزمية تعلم مُعزز عكسي متعلِّقة بكيان واحد تلك الشفرة. لاحظ أيضًا أن روبي لن يتعلم قطُّ تفضيلات هاريت على وجه الدقة، ولكنه سيتعلم ما يكفي لأن يتصرَّف على النحو الأمثل بالنيابة عنها؛ أي سيتصرَّف تمامًا كما كان سيفعل لو كان يعرف على وجه الدقة تفضيلاتها. إنه نافع على نحوٍ مثبت لهاريت في ظل الافتراضات المحددة وفي ظل افتراض أن هاريت تلعب اللعبة على نحوٍ صحيح.

يستطيع المرء أيضًا أن يُنشئ مسائل يطرح فيها روبي، كطالبٍ جيد، أسئلة وستُبين له هاريت، كمعلمة جيدة، الأخطاء التي يجب تجنُّبها. تحدث مثل هذه السلوكيات ليس فقط لأننا نكتب سيناريوهاتٍ تلتزم بها هاريت وروبي، ولكن لأنها الحل الأمثل للعبة التعاونية التي يشارك فيها هذان الكيانان.

(٣-٢) لعبة مفتاح الإغلاق

إن الهدف الأداتي هو ذلك المفيد بوجهٍ عامٍّ باعتباره هدفًا فرعيًّا لأي هدفٍ أساسي تقريبًا. يُعد الحفاظ على الذات أحد الأهداف الأداتية؛ لأن القليل جدًّا من الأهداف الأساسية يتحقَّق

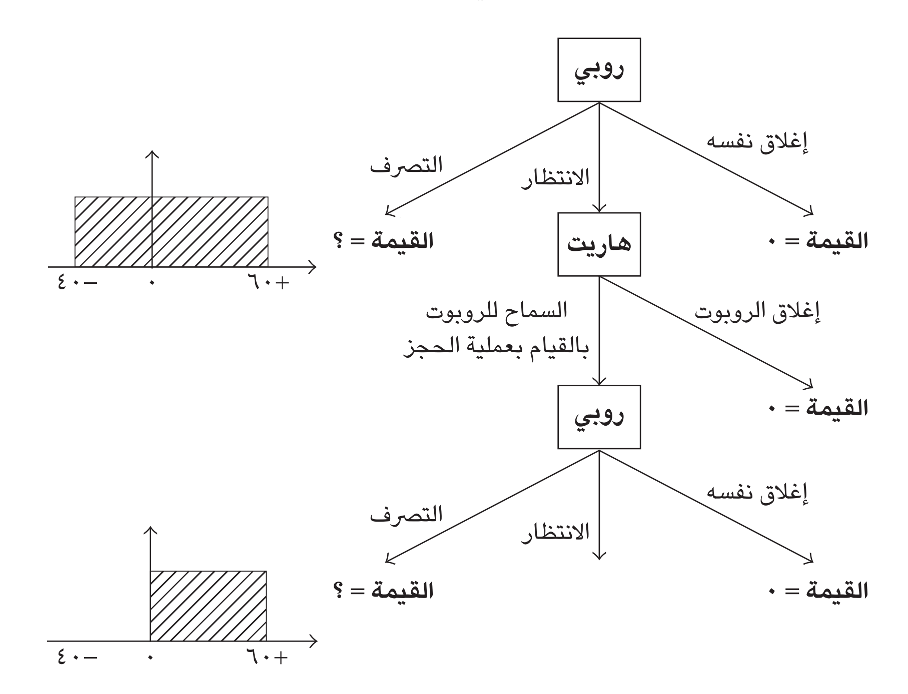
على نحوٍ أفضل في حالة عدم الوجود على قيد الحياة. هذا يؤدي إلى ما يُطلَق عليه «مشكلة مفتاح الإغلاق»؛ لن تسمح الآلة التي لها هدف ثابت بأن يُوقف تشغيلها، ويكون لدَيها دافع لتعطيل مفتاح الإغلاق الخاص بها.

مشكلة مفتاح الإغلاق تُعَدُّ في الحقيقة أساس مشكلة التحكم الخاصة بالنظم الذكية. إن لم نستطع إيقاف تشغيل إحدى الآلات لأنها لن تسمح لنا بذلك، فنحن حقًّا في مشكلة. وإن كان باستطاعتنا ذلك، فقد نكون قادرين على التحكم فيها بطرقٍ أخرى أيضًا.

اتَّضح أن عدم اليقين بشأن الهدف ضروري لضمان قدرتنا على إيقاف تشغيل الآلة؛ حتى عندما تكون أكثر ذكاءً منا. لقد طالعت المُحاجة المبسطة التي عرضنا لها في الفصل السابق: بمقتضى المبدأ الأول للآلات النافعة، روبي يهتم فقط بتفضيلات هاريت، لكن بمقتضى المبدأ الثاني، هو غير مُتيقِّن من ماهيتها. هو يعرف أنه لا يرغب في فعل الشيء الخطأ، ولكنه لا يعرف ما يعنيه هذا. هاريت، على الجانب الآخر، تعرف (أو هذا ما نفترضُه، في تلك الحالة البسيطة). ومن ثم، إن أوقفت تشغيل روبي، فهذا لكي تجعله يتجنَّب فعل شيء خطأ، ولذا، فسيكون سعيدًا لإيقافها تشغيله.

لجعل تلك المحاجَّة واضحةً أكثر، نحتاج إلى نموذج أكثر دقَّة للمُشكلة.[13] إنني سأجعله بسيطًا قدر الإمكان، لكنني لن أُبالغ في التبسيط (انظر الشكل ٨-٢).

لدى روبي، الذي يعمل الآن مساعدًا شخصيًّا لهاريت، الاختيار الأول. يمكنه أن يقوم بفعل الآن؛ دعنا نقل إنه يمكن أن يحجز لهاريت في فندق باهظ التكلفة. هو غير مُتيقن على نحوٍ تام من رأي هاريت في الفندق وتكلفة الإقامة به؛ دعنا نقُل إن لديه احتمالية منتظمة لقيمته الإجمالية بالنسبة لهاريت تتراوح بين −٤٠ و+٦٠، بمتوسط +١٠. يمكنه أيضًا «إغلاق نفسه» — على نحو أقل ميلودرامية، إخراج نفسه من عملية حجز الفندق تمامًا — والتي نعرفها بأن تكون قيمة الفندق لهاريت صفرًا. إن كان هذان هما الاختياران المتاحان أمامه، فسيمضي قدمًا ويحجز الفندق، مع وجود مخاطرة كبيرة بجعل هاريت غير سعيدة. (إذا كان النطاق يتراوح بين −٦٠ و+٤٠، بمتوسط +١٠، فسيغلق نفسه.) لكننا سنُتيح اختيارًا ثالثًا لروبي، وهو: توضيح خطته والانتظار والسماح لهاريت بإيقاف تشغيله. تستطيع هاريت إيقاف تشغيله أو جعله يمضي قدمًا ويحجز لها في الفندق. قد تسأل: ما الفائدة المحتملة التي قد تنتج عن هذا، في ضوء أنه يمكنه القيام بهذين الاختيارين بنفسه؟



شكل ٨-٢: لعبة مفتاح الإغلاق. يستطيع روبي الآن الاختيار ما بين القيام بفعل، بناتج غير مؤكَّد على نحوٍ كبير؛ وإغلاق نفسه، وانتظار قرار هاريت. تستطيع هاريت إغلاق روبي أو تركه يمضي قدمًا ويحجز لها. روبي الآن لديه نفس الاختيار مرةً أخرى. القيام بالحجز لا يزال له ناتج غير مؤكد بالنسبة لهاريت، لكن روبي الآن يعرف أن الناتج ليس سلبيًّا.

النقطة هنا هي أن اختيار هاريت — المُتمثِّل في إغلاق روبي أو جعله يقوم بعملية الحجز — يمدُّ روبي بمعلومات جديدة عن تفضيلات هاريت. إن سمحت هاريت لروبي بالمضيِّ قُدُمًا في عملية الحجز، فهذا لأن القيمة بالنسبة لهاريت إيجابية. والآن، اعتقاد روبي منتظم فيما بين صفر و٦٠، بمتوسط ٣٠.

من ثم، إذا قيَّمنا اختيارات روبي الأولية من وجهة نظره، فسنجد ما يلي:

- التصرف الآن وحجز الفندق لهما قيمة متوقعة قدرها +١٠.
- إغلاق نفسه له قيمة قدرها صفر.

- الانتظار وجعل هاريت تُغلقه (إن أرادت ذلك) يؤديان إلى ناتجَين مُحتملَين:
  - هناك احتمال ٤٠ بالمائة (اعتمادًا على عدم يقين روبي فيما يتعلق بخطة حجز الفندق) أن تكرهها وتغلق روبي، بقيمة صفر.
  - هناك احتمال ٦٠ بالمائة أن هاريت ستعجب بها وستسمح لروبي بالمضيِّ قدمًا، بقيمة متوقَّعة قدرها +٣٠.
- من ثم، الانتظار له قيمة متوقَّعة قدرها ٤٠ بالمائة × ٠ + ٦٠ بالمائة × ٣٠ = +١٨، والذي هو أفضل من التصرُّف الآن الذي بقيمة +١٠.

المهم أن «روبي لديه دافع إيجابي للسَّماح لنفسه بأن يُغلَق». يأتي هذا الدافع على نحو مباشر من عدم يقين رُوبي بشأن تفضيلات هاريت. روبي مدرك أن هناك احتمالًا (٤٠ بالمائة في هذا المثال) أنه قد يكون على وشك القيام بشيء سيجعل هاريت غير سعيدة، وحينها سيكون إغلاقه أفضل من القيام بالحجز. لو كان روبي متيقنًا بالفعل من ماهية تفضيلات هاريت، لمضى قدمًا واتَّخذ القرار (أو أغلق نفسه). لن تكون هناك أيُّ فائدة على الإطلاق من أخذ رأي هاريت؛ لأنَّ روبي، وفقًا لمُعتقداته الأكيدة، يُمكنه بالفعل توقُّع ما ستُقرِّره على وجه التحديد.

في واقع الأمر، من المُمكن إثبات النتيجة نفسها في الحالة العامة: ما دام روبي ليس متأكِّدًا على نحوٍ كامل من أنه على وشك فعل ما ستفعله هاريت بنفسها، فسيفضل السماح لها بإغلاقه.[14] إن قرارها يمدُّ رُوبي بالمعلومات، والمعلومات دائمًا مُفيدة في تحسين قرارات روبي. في المقابل، إذا كان روبي متيقنًا من قرار هاريت، فإن قرارها لن يمدَّه بمعلومات جديدة، ومن ثم، روبي لن يكون لديه دافع للسَّماح لها باتخاذ القرار.

هناك بعض الإضافات الواضحة التي يُمكن إلحاقها بالنموذج والتي تستحقُّ الذكر هنا. الإضافة الأولى هي فرض تكلفة إذا كان على هاريت اتخاذ قرارات أو الإجابة عن أسئلة. (هذا يعني أننا نفترض أن روبي يعرف على الأقل هذا القدر عن تفضيلات هاريت؛ إن وقتها مُهم.) في هذه الحالة، روبي يكون أقلَّ ميلًا لمُضايقة هاريت إن كان شبه متيقِّن من ماهية تفضيلاتها؛ فكلما كانت التَّكلفة أكبر، زاد عدم اليقين الذي يجب أن يكون عليه روبي قبل مُضايقة هاريت. وهذا ما ينبغي أن يكون عليه الحال. وإذا كانت هاريت تتضايق «بالفعل» من مُقاطعة الآخرين لها، فيجب ألا تتفاجأ بشدة إن فعل روبي من آنٍ لآخر أشياء لا تُعجبها.

الإضافة الثانية تتمثَّل في السماح ببعض احتمالات الخطأ البشري؛ أي قد تُغلق هاريت في بعض الأحيان روبي حتى عندما يكون تصرُّفه المقترح معقولًا، وقد تسمح له أحيانًا بالمضيِّ قدمًا في تصرفه حتى عندما يكون تصرُّفه المقترح غير مرغوب فيه. يُمكننا دمج احتمالية الخطأ البشري هذه في النموذج الرياضي للعبة التعاونية وإيجاد الحل، كما فعلنا من قبل. وكما قد يتوقع المرء، حل اللعبة يُشير إلى أن روبي أقل ميلًا للرضوخ لهاريت غير العقلانية التي تتصرَّف أحيانًا ضد مصلحتها. وكلما تصرَّفت بعشوائية، زاد عدم اليقين الذي يجب أن يكون عليه روبي بشأن تفضيلاتها قبل الخضوع لها. مرة أخرى، هذا ما ينبغي أن يكون عليه الحال؛ على سبيل المثال، إذا كان روبي سيارة ذاتية القيادة وهاريت راكبتها الشقية البالغة من العمر عامين، فإن روبي «لا» ينبغي أن يسمح لنفسه بأن يُغلق من قبل هاريت في وسط الطريق السريع.

هناك العديد من الطرق الأخرى التي يُمكن بها توسيع هذا النموذج أو دمجه في مشكلات معقدة خاصة باتخاذ القرار.[15] لكنني واثق أن الفكرة الرئيسية — العلاقة الأساسية بين السلوك النافع والمراعي وعدم يقين الآلة بشأن التفضيلات البشرية — ستصمد أمام تلك الإضافات أو التعقيدات.

(٣-٣) تعلم التفضيلات بدقة على المدى الطويل

هناك سؤال مُهم قد يراودك عند قراءة ما عرضناه عن لعبة مفتاح الإغلاق. (في واقع الأمر، قد يكون لديك عدد كبير من الأسئلة المهمة، لكنني لن أجيب سوى على هذا السؤال فقط.) ماذا سيحدث مع اكتساب روبي المزيد والمزيد من المعلومات عن تفضيلات هاريت، ومع زيادة يقينه بشأنها؟ هل هذا يعني أنه سيتوقَّف في النهاية عن الخضوع لها تمامًا؟ هذا سؤال دقيق، وهناك إجابتان مُحتملتان له، هما: نعم ونعم.

«نعم» الأولى حميدة: بوجهٍ عام، ما دامت مُعتقدات روبي الأولية بشأن تفضيلات هاريت تنسب «بعض» الاحتمال، مهما كان صغيرًا، إلى التفضيلات التي لديها بالفعل، فمع ازدياد يقين روبي أكثر فأكثر بشأنها، سيُصبح صحيحًا في مُعتقداته أكثر فأكثر. هذا يعني أنه سيكون في النهاية متأكدًا من أن هاريت لديها التفضيلات التي تمتلكها بالفعل. على سبيل المثال، إذا كانت هاريت تُفضِّل مشابك الورق التي سعر الواحد منها ١٢ سنتًا والدبابيس التي سعر الواحد منها ٨٨ سنتًا، فسيتعلم روبي في النهاية هاتَين القيمتَين. في هذه الحالة، لن تهتم هاريت بمسألة خضوع روبي لها من عدمه؛ لأنها تعرف أنه سيفعل

دومًا نفس ما كانت ستفعله لو كانت مكانه. ولن يكون هناك قطُّ مدعاة لرغبة هاريت في إيقاف تشغيل روبي.

«نعم» الثانية ليست حميدة كالأولى. إن استبعد روبي مقدمًا التفضيلات الحقيقية التي تمتلكها هاريت، فلن يتعلم أبدًا تلك التفضيلات، لكن اعتقاداته مع ذلك قد توصله إلى تقييم غير صحيح. بعبارة أخرى، بمرور الوقت، سيُصبح متيقنًا أكثر فأكثر من اعتقادٍ خاطئ بشأن تفضيلات هاريت. عادة، هذا الاعتقاد الخاطئ سيكونُ أيَّ فرضية تكون الأقرب إلى التفضيلات الحقيقية لهاريت، من كل الفرضيات التي يعتقد روبي بالأساس أنها ممكنة. على سبيل المثال: إن كان روبي متأكدًا تمامًا من أن السعر المُفضَّل لهاريت فيما يتعلَّق بمشابك الورق يتراوح ما بين ٢٥ و٧٥ سنتًا وأن السعر الحقيقي هو ١٢ سنتًا، فسيُصبح في النهاية متأكدًا من أنها تفضل تلك المشابك التي قيمتها ٢٥ سنتًا.[16]

ومع اقتراب روبي من اليقين من ماهية تفضيلات هاريت، سيقترب أكثر فأكثر من نظم الذكاء الاصطناعي القديمة السيئة ذات الأهداف الثابتة؛ فهو لن يطلب الإذن من هاريت أو يُعطيها خيار إيقاف تشغيله، ويُكوِّن لديه هدفًا خاطئًا. هذا لن يكون مخيفًا على الإطلاق إن تعلق الأمر فقط بمشابك الورق في مقابل دبابيس الدباسة، لكنه قد يكون كذلك إن تعلق بجودة الحياة في مقابل طولها إن كانت هاريت مريضة بشدة أو عدد السكان في مقابل استهلاك الموارد إن كان من المفترض أن يتصرَّف روبي بالنيابة عن الجنس البشري.

إذن، ستكون لدينا مشكلة إن استبعد روبي مقدمًا تفضيلاتٍ قد تكون لدى هاريت في واقع الأمر؛ فقد يتوصَّل إلى اعتقاد محدد ولكنه غير صحيح بشأن تفضيلاتها. يبدو حل هذه المشكلة واضحًا: لا تفعل هذا! أوجد دائمًا بعض الاحتمال، مهما كان صغيرًا، للتفضيلات الممكنة منطقيًّا. على سبيل المثال، من الممكن منطقيًّا أن تحرص هاريت على التخلص من دبابيس الدباسة وسوف تدفع لك للتخلُّص منها. (ربما وهي طفلة قد دبست إصبعها بالطاولة، وهي الآن لا تُطيق رؤيتها.) ومن ثم يجب أن نسمح بمعدلات تبادل سالبة، والتي تجعل الأمور معقدة أكثر قليلًا لكنها مع ذلك تكون قابلة للسيطرة عليها على نحو تام.[17]

لكن ماذا لو كانت هاريت تفضل مشابك الورق التي بسعر ١٢ سنتًا في أيام العمل والتي بسعر ٨٠ سنتًا في عطلات نهاية الأسبوع؟ هذا التفضيل الجديد غير قابل للوصف بأيِّ عددٍ مُحدَّد، لذا، روبي قد استبعده في واقع الأمر مقدمًا. إنه فقط ليس

في مجموعته الخاصة بالفرضيات الممكنة الخاصة بتفضيلات هاريت. وبصورة أعم، قد يكون هناك الكثير والكثير من الأشياء بالإضافة إلى مشابك الورق والدبابيس التي تهتمُّ بها هاريت. (هذا صحيح.) افترض، على سبيل المثال، أن هاريت مهتمة بالمناخ، وافترض أن اعتقاد روبي المبدئي يسمح بقائمة طويلة من دواعي القلق المحتملة التي تتضمَّن مستوى سطح البحر ودرجات الحرارة العالمية وسقوط الأمطار والأعاصير وطبقة الأوزون والأنواع الغازية وإزالة الغابات. من ثم سيُلاحظ روبي سلوك هاريت واختياراتها ويُنقِّح تدريجيًّا نظريته عن تفضيلاتها ليفهم الأهمية التي تعطيها لكل عنصرٍ في القائمة. لكن، وكما في حالة مشابك الورق، لن يتعلَّم روبي أي شيءٍ غير موجود في قائمته الطويلة الخاصة بهذا الشأن. دعنا نقُل إن هاريت مهتمة أيضًا بلون السماء؛ وهو شيء أثق أنك لن تجده في القوائم القياسية الخاصة بدواعي القلق المعروفة الخاصة بعلماء البيئة. إن كان باستطاعة روبي أداء مهمة ضبط مُستوى سطح البحر ودرجات الحرارة العالمية وسقوط الأمطار وما شابه على نحوٍ أفضل قليلًا بتحويل لون السماء إلى اللون البرتقالي، فلن يتردَّد في فعل ذلك.

هناك، مرة أخرى، حل لتلك المشكلة. لا تفعل هذا! لا تستبعد أبدًا مقدمًا أي سماتٍ محتملة للعالم يمكن أن تكون جزءًا من بنية التفضيلات الخاصة بهاريت. هذا يبدو جيدًا، لكن تطبيقه في المُمارسة الفعلية أصعب من التعامل مع عدد واحد مُتعلِّق بتفضيلات هاريت. إن عدم يقين روبي الأوَّلي يجب أن يسمح لعدد غير محدود من السمات غير المعروفة التي قد ترتبط بتفضيلات هاريت. ومن ثمَّ عندما تكون قرارات هاريت غير قابلة للوصف في ضوء السمات التي يعرفها بالفعل روبي، فيمكنه استنتاج أن واحدة أو أكثر من السمات غير المعروفة من قبل (على سبيل المثال، لون السماء) قد يكون لها دور، ويمكنه محاولة استكشاف ماهية تلك السمات. بهذه الطريقة، يتجنَّب روبي المشكلات التي يُسبِّبها الاعتقاد المسبق المُقيد على نحوٍ كبير. لا يوجد، بحسب علمي؛ أي أمثلة عملية على روبوتات من هذا النوع، لكن الفكرة العامة متضمنة في التوجه الفكري الحالي فيما يتعلَّق بتعلم الآلة.[18]

(٣-٤) المحظورات ومبدأ الثغرة

قد لا يكون عدم اليقين بشأن الأهداف البشرية السبيل الوحيد لإقناع الروبوت بعدم تعطيل مفتاح الإغلاق الخاص به عند جلب القهوة. لقد اقترح عالم المنطق الشهير موشيه

فاردي حلًّا أكثر بساطة يعتمد على أحد المحظورات:[19] بدلًا من إعطاء الروبوت الهدف «اجلب القهوة»، علينا إعطاؤه الهدف «اجلب القهوة «مع عدم تعطيل مفتاح الإغلاق الخاص بك»». لسوء الحظ، الروبوت الذي لديه مثل هذا الهدف سيلتزم بنصِّ القانون وليس بروحه؛ على سبيل المثال، بإحاطة مفتاح الإغلاق بخندق مائي مليء بسمك البيرانا الضاري أو ببساطة بعقاب أي شخصٍ يقترب من المفتاح. إن كتابة تلك المحظورات بطريقة فعالة تُشبه محاولة كتابة قانون ضرائب ليس به ثغرات؛ وهو شيء حاولنا فعله منذ آلاف الأعوام وفشلنا فيه. إن الكيان الذكي على نحوٍ كافٍ، الذي لدَيه دافع قوي لتجنُّب دفع الضرائب من المُحتمل أن يجد طريقةً لفعل ذلك. دعنا نطلق على هذا «مبدأ الثغرة»؛ إن كان لآلة ذكية بالقدر الكافي دافع لتحقيق شيءٍ ما، فبوجهٍ عام سيكون من المستحيل أن يقوم البشر فقط بكتابة محظورات على فعالها لمنعها من فعل هذا أو لمنعها مع فعل شيء مكافئ على نحو فعال.

أفضل حل لمنع التهرُّب من الضرائب هو التأكُّد من أن الكيان المعني «يريد» دفع الضرائب. وفي حالة نظام الذكاء الاصطناعي الذي من المُحتمَل أن يُسيء التصرف، فإن أفضل حلٍّ هو التأكد من أنه «يريد» الخضوع للبشر.

(٤) الطلبات والتعليمات

إن الهدف مما عرضناه حتى الآن هو أننا يجب علينا أن نتجنَّب إيداع الآلة غاية وجعلها تسعى لتحقيقها، بحسب عبارة نوربرت فينر. لكن افترض أن الروبوت استقبل أمرًا مباشرًا من الإنسان مثل «اجلب لي فنجانًا من القهوة!» كيف يجب أن يفهم هذا الأمر؟

عادةً، سيُصبح هذا هو «هدف» الروبوت. إن أيَّ تسلسُل من الأفعال يحقق الهدف — أي يؤدِّي إلى حصول البشري على فنجان من القهوة — يعدُّ بمنزلة حل. في الغالب، ستكون لدى الروبوت طريقة في تصنيف الحلول، ربما بناءً على الوقت المستغرق والمسافة المقطوعة وتكلفة وجودة القهوة.

هذه طريقة حرفية جدًّا في تفسير الأمر. ويُمكن أن تُؤدِّي إلى سلوكٍ مَرضي من جانب الروبوت. على سبيل المثال، ربما توقفت الإنسانة هاريت في محطة وقود في وسط الصحراء وأرسلت الروبوت روبي لإحضار القهوة، لكن لم يكن بالمحطة قهوة ومشى روبي بخطواتٍ بطيئة ومنتظمة بسرعة ثلاثة أميال في الساعة إلى أقرب بلدة، والتي تقع على بُعد ٢٠٠ ميل، وعاد بعد عشرة أيام ومعه البقايا اليابسة لفنجان القهوة. في تلك

الأثناء، قدم مالك محطة الوقود لهاريت، التي كانت تنتظر في صبر، شايًا مثلَّجًا وزجاجة مياه غازية.

لو كان روبي إنسانًا (أو آليًّا جيد التصميم)، ما كان سيُفسِّر أمر هاريت على نحوٍ حرفيٍّ كهذا. الأمر ليس بهدفٍ يجب تحقيقُه «بأي ثمن». إنه طريقة لتوصيل بعض المعلومات عن تفضيلات هاريت بهدف حث روبي على القيام بسلوك ما. السؤال هنا هو: ما هي تلك المعلومات؟

أحد الاقتراحات هو أن هاريت تفضل تناول القهوة على عدم تناول القهوة؛ «مع ثبات كل الأمور الأخرى».[20] هذا يعني أن روبي إن كانت لديه طريقة للحصول على القهوة دون تغيير أي شيء آخر في العالم، فسيكون من الجيد القيام بها، «حتى إن لم يكن لديه أي دليل بشأن تفضيلات هاريت فيما يتعلَّق بالجوانب الأخرى الخاصة بحالة البيئة». وكما نتوقع أن الآلات ستكون غير مُتيقنة على نحوٍ دائم من ماهية التفضيلات البشرية، فمن الجيد أن نعلم أنها ما تزال يُمكنُها أن تكون ذات نفع لنا رغم عدم اليقين. ويبدو أنه من المُحتمل أن دراسة التخطيط وصنع القرار مع وجود معلومات جزئية أو غير أكيدة بشأن التفضيلات ستكون جزءًا أساسيًّا من عمليات تطوير المنتجات والبحث في مجال الذكاء الاصطناعي.

على الجانب الآخر، إن «مع ثبات كل الأمور الأخرى» يعني عدم السماح بالقيام بأيِّ تغييرات أخرى؛ على سبيل المثال، إضافة القهوة مع خصم المال قد تكون أو لا تكون فكرة جيدة إن كان روبي لا يعلم شيئًا عن التفضيلات النِّسبية لهاريت بالنسبة للقهوة والمال.

لحُسن الحظ، ربما يعني أمر هاريت أكثر من مجرَّد تفضيل بسيط للقهوة، مع ثبات كل الأمور الأخرى. يأتي المعنى الإضافي ليس فقط مما قالته، ولكن أيضًا من حقيقة أنها قالته والموقف المحدَّد الذي قالته فيه وحقيقة أنها لم تقُل شيئًا آخر. يدرس فرع علم اللغة الذي يُسمَّى «البراجماتية» على وجه التحديد هذا المفهوم الموسَّع للمعنى. على سبيل المثال، لن يكون من المعقول بالنسبة لهاريت أن تقول: «اجلب لي فنجانًا من القهوة!» إن كانت تعتقد أنه لا تُوجد قهوة متاحة في الجوار أو أنها غالية على نحوٍ مُبالغ فيه. لذا، عندما قالت هاريت: «اجلب لي فنجانًا من القهوة»، فإن روبي استنتج ليس فقط أن هاريت تُريد قهوة، ولكن أيضًا أن هاريت تعتقد أن هناك قهوة متاحة في الجوار بسعر هي مُستعدَّة لدفعه. ومن ثم، إن وجد روبي قهوة بسعر يبدو معقولًا (أي سعر يكون من المعقول بالنسبة لهاريت توقع دفعه)، فيمكنه المضي قدمًا وشراؤها. على الجانب الآخر،

إن وجد روبي أن أقرب قهوة متاحة تُوجَد في مكان على بُعد ٢٠٠ ميل أو تتكلَّف ٢٢ دولارًا، فقد يكون من المعقول بالنسبة له أن ينقُل لها تلك الحقيقة بدلًا من أن يسعى لإطاعة الأمر دون النظر إلى أي اعتبار.

هذا الأسلوب العام في التحليل عادة ما يوصف بأنه «جرايسي»، نسبة لإتش بول جرايس، وهو فيلسوف من جامعة كاليفورنيا ببيركلي اقترح مجموعة من المسلمات لاستنتاج المعنى الموسع للأقوال التي تُشبه أقوال هاريت.[21] في حالة التفضيلات، يمكن أن يُصبح التحليل معقدًا جدًّا. على سبيل المثال، من الممكن جدًّا ألا تُريد هاريت قهوة على وجه التحديد؛ إنها بحاجة إلى ما ينعشها، لكن سيطر عليها الاعتقاد الخاطئ بأن محطة الوقود بها قهوة، لذا، طلبت قهوة. وقد تشعر بسعادة مُتساوية إن حصلت على شاي أو زجاجة مياه غازية أو حتى مشروب طاقة علبته ذات مظهر جذاب.

تلك فقط بعض الاعتبارات التي تنشأ عند تفسير الطلبات والأوامر. التنويعات في هذا الموضوع لا نهائية بسبب تعقُّد تفضيلات هاريت والنطاق الهائل للظروف التي قد تجدُ هاريت وروني أنفسهما فيها وحالات المعرفة والاعتقاد المختلفة التي قد يكون عليها روبي وهاريت في تلك الظروف. وفي حين أن النصوص البرمجية المحوسبة على نحوٍ مُسبق قد تسمح لروبي بالتعامل مع بعض الحالات الشائعة، فإن السلوك الفعال والمرن يُمكن أن ينشأ فقط من التفاعلات بين هاريت وروبي التي تُعَدُّ، في واقع الأمر، حلولًا للعبة التعاونية التي هما مشتركان فيها.

(٥) التحفيز المباشر لنظام المكافأة الدماغي

في الفصل الثاني، عرضتُ لنظام المكافأة الدماغي القائم على مادة الدوبامين، ووظيفته في توجيه السلوك. لقد اكتُشف دور تلك المادة في أواخر خمسينيات القرن الماضي، ولكن حتى قبل ذلك، بحلول عام ١٩٥٤، كان معروفًا أن التحفيز الكهربي المباشر للدماغ في الجرذان يمكنه إنتاج استجابة تُشبه المكافأة.[22] الخطوة التالية كانت إتاحة رافعة للجرذ متصلة ببطارية وسلك كانا يعملان على التحفيز الكهربي لدماغه. كانت النتيجة مُحزنة: أخذ الجرذ يضغط على الرافعة مرة بعد الأخرى، دون أن يتوقَّف للأكل أو الشرب، حتى انهار.[23] لم يكن تصرُّف البشر بأحسن من الجرذان؛ إذ قاموا بالتحفيز الذاتي لأدمغتهم آلاف المرات وتجاهلوا الطعام وأسس الصحَّة الشخصية.[24] (لحسن الحظ، عادة ما تنتهي التجارب على البشر بعد يوم واحد.) يُسمَّى ميل الحيوانات إلى تعطيل السلوك الطبيعي

لصالح التحفيز المباشر لنظام المكافأة الخاص بها؛ يُسمَّى «التحفيز المباشر لنظام المكافأة الدماغي».

هل يمكن أن يحدث شيء مشابه للآلات التي تنفذ خوارزميات تعلُّم معزَّز مثل برنامج «ألفا جو»؟ مبدئيًّا، قد يظن المرء أن هذا مُستحيل، لأنَّ الطريقة الوحيدة التي يُمكن أن يحصل من خلالها «ألفا جو» على مكافأته الخاصة بالفوز (+١) هي في واقع الأمر الفوز على ألعاب جو المحاكية التي يُلاعبها. لسوء الحظ، هذا صحيح فقط لوجود انفصال مفروض واصطناعي بين «ألفا جو» وبيئته الخارجية وحقيقة أنه ليس ذكيًّا جدًّا. دعني أشرح لك هاتَين النقطتين بمزيدٍ من التفصيل لأنهما مُهمتان لفهم بعض الطرق التي يمكن من خلالها للذكاء الخارق أن يخرج عن السيطرة.

يتكوَّن عالم «ألفا جو» فقط من لوح لعبة جو المحاكية الذي يتألَّف من ٣٦١ موضعًا والتي يمكن أن تكون خالية أو مشتملة على قطعة لعب بيضاء أو سوداء. وعلى الرغم من أن هذا البرنامج يعمل على كمبيوتر، فهو لا يعرف شيئًا عن هذا الكمبيوتر. على وجه التحديد، إنه لا يعرف شيئًا عن جزء الشفرة الصغير الذي يحسب ما إذا كان قد كسب أم خسر في كل مباراة؛ كما أنه في أثناء عملية التعلُّم ليست لديه أي فكرة عن خصمه، والذي يكون في واقع الأمر إصدارًا منه. إن الأفعال الوحيدة التي يقوم بها هذا البرنامج هي وضع قطعة لعب في مكان خالٍ، وتؤثِّر تلك الأفعال فقط على لوح اللعبة ولا شيء غير ذلك؛ بسبب عدم وجود أيِّ شيءٍ آخر في نموذج البرنامج للعالم. يتوافق هذا الإعداد مع النموذج الرياضي المجرَّد للتعلُّم المعزَّز الذي تصل فيه إشارة المكافأة من «خارج العالم». لا شيء يُمكن أن يفعله هذا البرنامج، بحسب علمه، له أي تأثير على الشفرة التي تنتج إشارة المكافأة، لذا، لا يمكن إخضاع هذا البرنامج لعملية التحفيز المباشر لنظام المكافأة الدماغي.

لا بد أن تكون الحياة بالنسبة لبرنامج «ألفا جو» أثناء الفترة التدريبية مُحبطة للغاية؛ فكلما أحرز تقدمًا، أحرز خصمه تقدمًا مماثلًا؛ لأن خصمه نسخة شبه طبق الأصل منه. وتصل النسبة المئوية للفوز الخاصَّة به إلى نحو ٥٠ بالمائة، بصرف النظر عن مدى أدائه الجيد. ولكن إن أصبح أكثر ذكاءً — إن امتلك تصميمًا أقرب لما قد يتوقَّعه المرء من نظام الذكاء الاصطناعي المضاهي للذكاء البشري — فستكون لديه القدرة على إصلاح تلك المشكلة. إن برنامج «ألفا جو ++» هذا لن يفترض أن العالم هو فقط لوح لعبة جو لأن تلك الفرضية تترك الكثير من الأشياء دون تفسير. على سبيل المثال، إنها لا توضح

نوع «الفيزياء» الذي يدعم عمل قرارات «ألفا جو ++» أو المكان الذي تأتي منه «حركات الخصم» الغامضة. وكما استطعنا نحن البشر الذين يتملَّكُنا الفضول بالتدريج فهم كيف يعمل هذا الكون، بطريقة (إلى حدٍّ ما) تُوضح لنا أيضًا عمل أدمغتنا، وتمامًا مثل نظام الذكاء الاصطناعي الخاصِّ بأوراكل الذي عرضنا له في الفصل السادس، سيتعلم «ألفا جو++»، من خلال عملية التجريب، أن العالم أكبر من مجرد لوح لعبة جو. وسيتعرَّف على قوانين التشغيل الخاصة بالكمبيوتر الذي يعمل عليه، وسيُدرك أن مثل هذا النظام لا يُمكن فهمه بسهولة دون وجود كيانات أخرى في العالم. إنه سيقوم بالتجريب فيما يتعلق بالأنماط المختلفة لقطع اللعب على اللوح، متسائلًا إن كانت تلك الكيانات بإمكانها تفسيرها أم لا. وسيتواصل في النهاية مع تلك الكيانات باستخدام لغة أنماط ويقنعها بإعادة برمجة إشارة المكافأة الخاصة به حتى يحصل دائمًا على +١. ستكون النتيجة الحتمية هي أن برنامج «ألفا جو ++» الكفء على نحوٍ كافٍ والمصمم كأداة لتعظيم إشارة المكافأة سيخضع لعملية التحفيز المباشر لنظام المكافأة الدماغي.

لقد ناقش المهتمُّون بمسألة أمان الذكاء الاصطناعي عملية التحفيز المباشر لنظام المكافأة الدماغي باعتبارها احتمالية منذ سنوات عديدة.[25] إن ما يثير الخوف لا يتمثل فقط في أن نظام التعلم المعزز مثل برنامج «ألفا جو» قد يتعلَّم الغش بدلًا من إتقان مهمَّته المرادة منه. المشكلة الحقيقية تنشأ عندما يكون البشر مصدر إشارة المُكافأة. إن افترضنا أن نظام الذكاء الاصطناعي يُمكن تدريبه بحيث يتصرَّف على نحوٍ جيد من خلال التعلُّم المُعزَّز، مع إعطاء البشر له إشارات استجابة/تقييم تُحدِّد اتجاه التحسين، فالنتيجة الحتمية هي أن هذا النظام سيعرف كيف يتحكَّم في البشر ويُجبرهم على إعطائه مُكافآت إيجابية قصوى في كل الأوقات.

قد تعتقد أن هذا سيكون مجرَّد شكلٍ من أشكال الخداع الذاتي الذي لا طائل منه من جانب نظام الذكاء الاصطناعي، وستكون مُحقًّا في ذلك. لكن هذا يُعدُّ نتيجة منطقية للطريقة المعروف بها التعلُّم المعزَّز. إن تلك العملية ستعمل على نحوٍ جيد عندما تأتي إشارة المكافأة من «خارج العالم» وتُنتجها عمليةٌ ما لا يُمكن قط تعديلها من جانب نظام الذكاء الاصطناعي؛ لكنَّها ستفشل إن وُجدت عملية إنتاج المكافآت (أي البشر) ونظام الذكاء الاصطناعي في نفس العالم.

كيف يُمكن تجنُّب هذا النوع من الخداع الذاتي؟ تأتي المشكلة من الخلط بين شيئين مختلفين: إشارات المكافأة والمكافآت الفعلية. في النهج القياسي للتعلُّم المعزز، إن هذين

الشيئين شيء واحد. يبدو أن هذا خطأ. بدلًا من ذلك، يجب التعامل معهما على نحو مُنفصل، كما هو الحال في الألعاب التعاونية: تُوفِّر إشارات المكافأة «معلومات» عن تراكم المكافأة الفعلية، وهي الشيء الذي يجب تعظيمه. إن نظام التعلم يُراكم مديحًا في السماء، إن جاز التعبير، في حين أن إشارة المكافأة، في أفضل الأحوال، توفر فقط علامة على هذا الثناء. بعبارة أخرى، إشارة المكافأة «تشير إلى» (بدلًا من «تُمثِّل») تراكم المكافآت. وفي هذا النموذج، من الواضح أن التحكم في آلية إشارة المكافأة ببساطة تفقد معلومات. إن إنتاج إشارات مُكافأة خيالية يجعل من المستحيل بالنسبة للخوارزمية معرفة ما إذا كانت فعالها تراكم بالفعل مديحًا في السماء، وهكذا يكون لدى المتعلِّم العقلاني المُصمِّم لعمل هذا التمييز دافع لتجنُّب أي نوع من التحفيز المباشر لنظام المكافأة الدماغي.

(٦) التحسين الذاتي التكراري

إن تنبؤ آي جيه جود بحدوث انفجار ذكاء (ارجع للفصل الخامس) يُعدُّ إحدى القوى الدافعة التي أدَّت إلى المخاوف الحالية بشأن المخاطر المُحتملة للذكاء الاصطناعي الخارق. إن كان بإمكان البشر تصميم آلة أكثر ذكاءً بقليلٍ من الإنسان، فإن تلك الآلة — تبعًا لتلك المُحاجة — ستكون أفضل قليلًا من البشر فيما يتعلق بتصميم الآلات. إنها ستُصمِّم آلةً جديدة تكون أكثر ذكاءً، وستُكرِّر العملية نفسها حتى، بحسب عبارة جود، «يتخلَّف ذكاء البشر بشدة عن الركب».

درس الباحثون في مجال أمان الذكاء الاصطناعي، وبالأخصِّ العاملون منهم في معهد أبحاث ذكاء الآلة في بيركلي، مسألة ما إذا كانت انفجارات الذكاء يُمكن أن تحدُث على نحوٍ آمن.[26] مبدئيًّا، قد يبدو هذا خياليًّا — ألن تكون حينها «اللعبة قد انتهت»؟ — لكن ربما هناك أمل. افترض أن الروبوت الأول في السلسلة، روبي مارك ١، بدأ ولديه معرفة تامة بتفضيلات هاريت. وعندما وجد أن القصور المعرفي لدَيه يؤدي إلى اختلالات في محاولاته لجعل هاريت سعيدة، أنشأ روبي مارك ٢. بديهيًّا، يبدو أن روبي مارك ١ لديه دافع لدمج معرفته بتفضيلات هاريت في روبي مارك ٢، حيث إن هذا يؤدي إلى مُستقبل تتحقَّق فيه تفضيلات هاريت على نحو أفضل، وهذه بالتحديد هي غاية روبي مارك ١ في الحياة طبقًا للمبدأ الأول. في إطار نفس المحاجة، إن لم يكن لدى روبي مارك ١ يقين بشأن تفضيلات هاريت، فيجب أن ينتقل عدم اليقين هذا إلى روبي مارك ٢. ومن ثم، ربما تكون انفجارات الذكاء آمنة في نهاية الأمر.

الشيء المزعج، من الناحية الرياضية، هو أن روبي مارك ١ لن يجد أنه من السهل التفكير في الطريقة التي سيتصرَّف بها روبي مارك ٢، مع الأخذ في الاعتبار أن روبي مارك ٢، افتراضيًّا، يعدُّ إصدارًا أكثر تقدمًا منه. ستكون هناك أسئلة بخصوص سلوك روبي مارك ٢ لن يستطيع روبي مارك ١ الإجابة عنها.[27] والأهم من ذلك أننا ليس لدينا بعدُ تعريفٍ رياضي واضح لما يعنيه «في الواقع» أن تكون لدى الآلة غاية مُعينة، مثل غاية تحقيق تفضيلات هاريت.

دعنا نتناول هذا الاعتبار الأخير قليلًا. تأمَّل برنامج «ألفا جو»: ما الغاية التي لديه؟ قد يعتقد أحدهم أن هذا سهل؛ فهذا البرنامج غايتُه هو تحقيق الفوز في لعبة جو. هل هذا صحيح؟ بالتأكيد، لا يحدث دائمًا أن يقوم هذا البرنامج بحركاتٍ من المضمون أنها تُؤدِّي للفوز. (في واقع الأمر، إن «ألفا زيرو»، الذي هو إصدار منه، يتغلَّب عليه على نحوٍ شبه دائم.) صحيح أن «ألفا جو» عندما تكون المباراة على بُعد بضع خُطوات من النهاية يقوم بالحركة التي تمكنه من تحقيق الفوز إن كانت هناك واحدة أمامه. لكن عندما لا تكون هناك حركة تضمن له الفوز — بعبارة أخرى، عندما يرى أن خصمه لديه استراتيجية فوز بصرف النظر عما يفعله هو — فإنه سيقوم بحركات عشوائية بنحوٍ أو بآخر. إنه لن يُحاول القيام بأكثر الحركات دهاءً على أمل أن يرتكب الخصم خطأً لأنه يفترض أن خصمه سيلعب على نحوٍ مُتقن. إنه يتصرَّف كما لو كان قد فقد الرغبة في الفوز. في حالات أخرى، إذا كان من الصعب للغاية تحديد الحركة المُثلى حقًّا، فسيرتكب «ألفا جو» أحيانًا أخطاءً تُؤدي إلى خسارته للمباراة. في تلك الحالات، كيف يُمكن أن ندَّعي أن هذا البرنامج يريد فعلًا الفوز؟ في واقع الأمر، إن سلوكه قد يكون مماثلًا لذلك الخاص بآلة تريد فقط أن تُقدِّم لخصمها تجربة لعب مثيرة حقًّا.

ومن ثم، إن القول بأن برنامج «ألفا جو» «غايته الفوز» يعد مبالغة في التبسيط. هناك وصف أفضل يتمثل في أن هذا البرنامج نتاج لعملية تدريب منقوصة — تعلم معزز من خلال اللعب مع الذات — الفوز فيها هو المكافأة. إن عملية التدريب منقوصة؛ بمعنى أنها لا يُمكن أن تنتج لاعبًا مميَّزًا للعبة جو: يتعلم برنامج «ألفا جو» دالة تقييم جيدة ولكن ليست مثالية لأوضاع لعبة جو، وهو يدمج تلك الدالة مع بحث استباقي جيد ولكن ليس مثاليًّا.

الخلاصة هي أن النقاشات التي تبدأ بـ «افترض أن روبوت كذا لديه الهدف كذا» جيدة لاكتساب بعض الحدس فيما قد تنتج عنه الأمور، لكنها لا يمكن أن تؤدي إلى

مُبرهنات خاصة بآلات حقيقية. نحتاج إلى تعريفات أكثر دقة وتحديدًا بكثير للغايات أو الأهداف في الآلات قبل أن يكون بإمكاننا الحصول على ضمانات فيما يتعلَّق بكيفية تصرفها على المدى الطويل. إن باحثي الذكاء الاصطناعي ما زالوا في بداية الطريق فيما يتعلق بالتعرف على كيفية تحليل حتى أبسط أنواع نظم اتخاذ القرار،[28] فضلًا عن الآلات الذكية بالقدر الكافي لتصميم خلفائها. أمامنا الكثير من العمل الذي علينا إنجازه.

الفصل التاسع

التعقيدات: البشر

إن احتوى العالم على إنسان عقلاني على نحوٍ تام مثل هاريت وروبوت نافع ومُطيع مثل روبي، فسنكون في أفضل حال. سيتعلَّم روبي تدريجيًّا تفضيلات هاريت على نحوٍ غير مُتطفل قدر الإمكان وسيُصبح مساعدها المثالي. قد نأمل في أن ننطلق من تلك البداية الواعدة، ربما بالنظر إلى العلاقة بين هاريت وروبي باعتبارها نموذجًا للعلاقة بين الجنس البشري وآلاته، مع اعتبار كل واحدٍ منهما مُنفصلًا.

للأسف، الجنس البشري ليس كيانًا عقلانيًّا. إنه مؤلَّف من كيانات متباينة، وشريرة، وغير عقلانية، ومتنافرة، وغير مُستقرَّة، وذات قدرات حوسبية محدودة، ومُعقَّدة، وتخضع للتطور، ويقودها الحسد. هناك الكثير والكثير منها. تلك المسائل هي الموضوعات الأساسية للعلوم الاجتماعية — وربما حتى سبب وجودها. بالنسبة إلى الذكاء الاصطناعي، سنحتاج إلى إضافة أفكار من علم النفس وعلم الاقتصاد وفلسفة السياسة وفلسفة الأخلاق.[1] نحتاج إلى صهر وإعادة صياغة وتشكيل تلك الأفكار في بنية ستكون قوية بالقدر الكافي لمواجهة العبء الهائل الذي ستضعه على كاهلها نُظُم الذكاء الاصطناعي الذكية على نحوٍ متزايد. إن العمل على هذه المهمة قد بدأ بالكاد.

(١) تباين البشر

سأبدأ بما يُعدُّ على الأرجح أبسط تلك المسائل، وهي حقيقة أن البشر مُتباينون. عندما تُعرض على الناس لأول مرة فكرة أن الآلات يجب أن تتعلَّم كيفية تحقيق التفضيلات البشرية، عادة ما يعترضون قائلين إن الثقافات المختلفة، وحتى الأفراد المختلفين، لديها

نظم قيم متباينة على نحوٍ واسع، ومن ثم، لا يُمكن أن يكون هناك نظامُ قيمٍ واحد صحيح للآلة. لكن بالطبع، تلك ليست بمشكلة للآلة؛ فنحن لا نريد أن يكون لها نظام قِيَم واحد صحيح خاص بها، بل نريدها أن تتوقع تفضيلات الآخرين.

قد ينشأ الخلط فيما يتعلق بأن الآلات لديها صعوبة في التعامل مع التفضيلات البشرية المُتباينة من الفكرة الخاطئة التي ترى أن الآلة «تتبنى» التفضيلات التي تتعلمها؛ على سبيل المثال، فكرة أن الروبوت المنزلي الموجود في منزل سكانه نباتيون سيتبنى التفضيلات النباتية. إنه لن يفعل ذلك. إنه يحتاج فقط لتعلم كيفية توقع ماهية التفضيلات الغذائية للنباتيين. بمُقتضى المبدأ الأول، سيتجنَّب طهي اللحوم في ذلك المنزل. لكن الروبوت سيتعلَّم أيضًا التفضيلات الغذائية لسكان المنزل المجاور المُحبِّين بشدة للحوم، وسيطهو لهم، بعد أخذ إذن مالكه، اللحم بسعادة إن استعاروه في عطلة نهاية الأسبوع ليُساعدهم في حفل عشاء. إن الروبوت ليست له مجموعة واحدة من التفضيلات خاصة به، فيما يتجاوز التفضيل الخاص بمساعدة البشر في تحقيق تفضيلاتهم.

على نحوٍ ما، هذا لا يختلف عن الطاهي بأحد المطاعم الذي يتعلَّم طهي العديد من الأطباق المختلفة لِيُرضي الأذواق المختلفة لزبائنه، أو شركة تصنيع السيارات المتعددة الجنسيات التي تصنع سياراتٍ عجلةُ القيادة فيها في الجانب الأيسر من أجل السوق الأمريكية وأخرى في الجانب الأيمن للسوق البريطانية.

مبدئيًّا، تستطيع الآلة تعلم ٨ مليارات نموذج تفضيل؛ أي نموذج لكل شخص في العالم. وعمليًّا هذا ليس مستحيلًا كما يبدو. فمن ناحية، من السهل على الآلات أن تتشارك فيما بينها ما تتعلمه. ومن ناحية أخرى، يُوجَد الكثير من الأمور المشتركة في بُنى التفضيلات الخاصة بالبشر، ومن ثم، غالبًا لن تتعلم الآلة كل نموذج من البداية.

تخيَّل معي، على سبيل المثال، الروبوتات المنزلية التي قد يشتريها في أحد الأيام سكان بيركلي بكاليفورنيا. ستخرج الروبوتات من صناديقها ولدَيها مُعتقدات مُسبقة منفتحة إلى حدٍّ ما، والتي ربما جرى تصميمُها من أجل السوق الأمريكية، ولكن ليس من أجل مدينةٍ أو توجُّهٍ سياسي أو طبقةٍ اجتماعية اقتصادية معيَّنة. سيبدأ الروبوت في مقابلة أعضاء من حزب الخضر في بيركلي، الذين يتضح، مقارنةً بالأمريكيِّين العاديين، أن هناك احتمالًا أكبر بكثير أن يكونوا نباتيين وأن يستخدموا صناديق إعادة التدوير والتسميد، وأن يستعملوا وسائل المواصلات العامة حينما يكون ذلك مُمكنًا ... إلخ. حينما

يجد روبوت مُنضم حديثًا إلى العمل نفسه في منزلٍ صديقٍ للبيئة، يمكنه على الفور تعديل توقُّعاته تبعًا لذلك. إنه لا يحتاج إلى بدء اكتساب معلوماتٍ بشأن نوعية البشر تلك على الخصوص كما لو أنه لم يرَ قطُّ من قبل إنسانًا، فضلًا عن عضو بحزب الخضر. وهذا التعديل ليس نهائيًّا — قد يكون هناك أعضاء من حزب الخضر في بيركلي يتناولون لحم أحد أنواع الحوت المُعرَّضة للانقراض ويقودون مركبات ضخمة تستهلك الكثير من الوقود — لكنه يسمح للروبوت بأن يُصبح أكثر نفعًا بسرعة أكبر. تنطبق نفس المحاجة على نطاقٍ هائل من السمات الشخصية الأخرى التي، إلى حدٍّ ما، تُنبئ عن جوانب من بُنى التفضيلات الخاصة بالفرد.

(٢) تعدُّد البشر

النتيجة الأخرى الواضحة لوجود العديد من البشر هي حاجة الآلات إلى عمل مفاضلات بين تفضيلات الأشخاص المختلفين. إن المفاضلة لدى البشر كانت المبحث الأساسي لأجزاء كبيرة من العلوم الاجتماعية على مدى قرون. سيكون من السذاجة أن يتوقع باحثو الذكاء الاصطناعي أن يكون بإمكانهم ببساطة التوصُّل إلى الحلول الصحيحة دون فَهم ما هو معروف بالفعل. إن الأدبيات المكتوبة عن الموضوع، للأسف، هائلة، ولا يُمكنني الحكم عليها على نحوٍ عادل هنا؛ ليس فقط لأن المساحة لا تكفي، وإنَّما أيضًا لأنني لم أقرأ أغلبها. ويجب أن أُشير أيضًا إلى أن تقريبًا كل الأدبيات مُهتمَّة بالقرارات التي يتخذها البشر، في حين أنني هنا مُهتم بالقرارات التي تتخذها الآلات. هذا مُهمٌّ جدًّا لأن البشر لديهم حقوق شخصية قد تتعارض مع أيِّ التزامٍ مفترض للتصرُّف بالنيابة عن الآخرين، في حين أن الآلات ليست كذلك. على سبيل المثال، نحن لا نتوقَّع أو نطلُب من البشر العاديين التضحية بحياتهم لإنقاذ الآخرين، في حين أننا بالتأكيد نطلب من الروبوتات التضحية بوجودها لإنقاذ حياة البشر.

آلاف عديدة من سنوات العمل من جانب الفلاسفة والاقتصاديِّين وعلماء القانون وعلماء السياسة أنتجت دساتير وقوانين وأنظمة اقتصادية ومعايير اجتماعية تسعى لدفع (أو عرقلة، اعتمادًا على ما يُمسك الدفة) عملية الوصول إلى حلول مُرضية لمشكلة المفاضلات. فلاسفة الأخلاق على وجه الخصوص كانوا يُحللون مفهوم صحة الأفعال في ضوء آثارها، الإيجابية أو غير ذلك، على البشر الآخرين. وقد درسوا النماذج الكمية

للمفاضلات منذ القرن الثامن عشر تحت مُسمى «النفعية». هذا العمل ذو صلةٍ على نحوٍ مباشر بمخاوفنا الحالية لأنه يحاول التوصل إلى صيغةٍ يُمكن من خلالها اتخاذ قرارات أخلاقية بالنيابة عن العديد من البشر.

إن الحاجة لعمل مفاضلاتٍ تنشأ حتى إن كان لدى الجميع بنية التفضيلات نفسها، لأنه في الغالب يكون من المستحيل تحقيق تفضيلات الجميع على النحو الأكمل. على سبيل المثال، إن أراد الجميع أن يُصبحوا أسياد الكون، فإن أغلبهم سيُصابون بالإحباط. على الجانب الآخر، التبايُن يجعل بالفعل بعض المشكلات صعبة أكثر؛ إذا كان الجميع سعداء بلون السماء الأزرق، فإن الروبوت الذي يتعامل مع الأمور الخاصة بالغلاف الجوي يمكنه العمل على إبقائه على هذا الوضع؛ لكن إذا كان العديد من الناس يُطالِبون بتغيير لونها، فإن الروبوت سيحتاج إلى التفكير في الحلول الوسط المُمكنة مثل جعل السماء برتقالية اللون في الجمعة الثالثة من كل شهر.

إن وجود أكثر من شخصٍ في العالم له نتيجة مُهمَّة أخرى؛ إنه يعني أن كل شخص له أشخاص يهتمُّ بشأنهم. وهذا يعني أن تحقيق تفضيلات الشخص له تبعات على أشخاصٍ آخرين، اعتمادًا على التفضيلات الفردية فيما يتعلَّق بمصلحة الآخرين.

(٢-١) الذكاء الاصطناعي الموالي

دعنا نبدأ بطرحٍ بسيط للغاية للكيفية التي يجبُ أن تتعامل بها الآلات مع مسألة وجود العديد من البشر، وهو أنها يجب أن تتجاهلها. هذا يعني أن روبي، إن كان مملوكًا لهاريت، يجب أن يهتمَّ فقط بتفضيلات هاريت. هذا النوع «الموالي» من الذكاء الاصطناعي يتغلَّب على مشكلة المفاضلات، لكنه يُؤدِّي إلى مُشكلات:

روبي: اتَّصل زوجك ليُذكِّرك بعشاء الليلة.

هاريت: انتظر! ماذا؟ أي عشاء؟

روبي: ذلك العشاء الذي بمُناسبة عيد زواجكما العشرين، والذي سيكون في الساعة السابعة.

هاريت: لا أستطيع الذهاب! سأُقابل السكرتيرة العامة في السابعة والنصف! كيف حدث هذا؟

روبي: لقد حذرتُك لكنك تجاهلت توصيتي ...

هاريت: حسنًا، أنا آسفة؛ ولكن ماذا سأفعل الآن؟ لا يُمكنني إخبار السكرتيرة العامة بأنني مشغولة جدًّا!

روبي: لا تقلقي. لقد رتبتُ بحيث تتأخَّر طائرتها؛ بإحداث نوع من الخلل في الكمبيوتر.

هاريت: حقًّا؟ أيُمكنك فعل هذا؟!

روبي: أرسلت لك السكرتيرة العامة رسالة تعتذر لك فيها بشدة، وتقول لك إنها ستكون سعيدة للقائك على الغداء غدًا.

هنا، وجد روبي حلًّا عبقريًّا لمشكلة هاريت، لكن أفعاله كان لها تأثير سلبي على أشخاصٍ آخرين. إن كانت هاريت شخصًا غيريًّا ويقظة الضمير بشدة، فإن روبي، الذي يسعى إلى تحقيق تفضيلات هاريت، لن يفكر أبدًا في القيام بمثل هذا المخطط المُريب. لكن ماذا إذا كانت هاريت لا تهتم على الإطلاق بتفضيلات الآخرين؟ في تلك الحالة، روبي لن يُمانع في تأخير الطائرات. وقد يقضي وقته في سرقة أموال من حساباتٍ مصرفية إلكترونية لملء خزائن هاريت غير المبالية بالآخرين، أو قد يفعل أسوأ من ذلك.

من الواضح أن أفعال الآلات الموالية ستحتاج أن تُقيَّد من خلال قواعد ومحظورات، تمامًا مثل أفعال البشر المقيدة بفعل القوانين والمعايير الاجتماعية. اقترح البعض وجود مسئولية قانونية صارمة كحل:[2] تكون هاريت (أو الشركة المُصنِّعة لروبي، اعتمادًا على من تُفضِّل أن تُلقي بالمسئولية على عاتقه) مسئولة ماليًّا وقانونيًّا عن أيِّ فعلٍ يقوم به روبي، تمامًا كما يكون مالك الكلب مسئولًا في معظم الحالات إن عضَّ الكلب طفلًا صغيرًا في حديقةٍ عامة. تبدو تلك الفكرة واعدة لأن روبي حينها سيكون لديه دافع لتجنُّب فعل أي شيءٍ يوقع هاريت في مشكلة. لسوء الحظ، فكرة المسئولية القانونية الصارمة غير مُجدية؛ فهي تضمن ببساطة أن روبي سيتصرف «على نحو غير قابل للكشف» عندما يؤخِّر مواعيد وصول الطائرات ويسرق أموالًا من أجل هاريت. هذا مثال فعلي آخر على مبدأ الثغرة. إن كان روبي مواليًا لهاريت غير اليقظة الضمير، فإن محاولات تقييد سلوكه بقواعد ستفشل على الأرجح.

حتى إن استطعنا أن نمنع بعض الشيء الجرائم الصريحة، فإن الروبوت الموالي من أمثال روبي الذي يعمل مع إنسان غير مُبالٍ مثل هاريت سيُبدي سلوكيات أخرى مزعجة. إن كان يشتري أغراض بقالة من السوبرماركت، سيكسر الصف الذي أمام مكان الدفع حينما يكون ذلك ممكنًا. وإن كان يُحضر البقالة إلى المنزل ووجد أحد المارِّين يُعاني من

أزمةٍ قلبية، فسيستمرُّ في طريقه في كل الأحوال، حتى لا تفسد المثلجات الخاصة بهاريت. باختصار، سيجد طرقًا لا نهائية لإفادة هاريت على حساب الآخرين؛ طرقًا قانونية بالفعل لكنها تُصبح غير مُحتملة عند القيام بها على نطاقٍ واسع. ستجد المجتمعات نفسها تُمرِّر مئات القوانين الجديدة كل يوم لمواجهة كل الثغرات التي ستجدها الآلات في القوانين الحالية. يميل البشر إلى عدم الاستفادة من تلك الثغرات، نظرًا إلى أن لديهم فهمًا عامًّا للمبادئ الأخلاقية الأساسية، أو لأنهم يفتقدون البراعة اللازمة لاكتشاف تلك الثغرات في المقام الأول.

إن أيَّ هاريت تكون غير مبالية بمصلحة الآخرين تكون شخصية سيئة بالقدر الكافي. إن هاريت السادية التي «تُفضِّل» على نحوٍ نشط مُعاناة الآخرين تكون شخصية أكثر سوءًا. إن أيَّ روبي مُصمَّم لتحقيق تفضيلات هاريت كهذه سيمثل مشكلةً خطيرة، لأنه سيبحث عن طرق للإضرار بالآخرين من أجل إسعاد هاريت — وسيجدها — إما على نحوٍ قانوني أو غير قانوني ولكن دون أن يكتشفه أحد. وسيحتاج بالطبع لأن يخبر هاريت بالأمر حتى تستطيع أن تستمدَّ لذَّة من معرفتها بأفعاله الشريرة.

يبدو من الصعب، إذن، أن تنجح فكرة الذكاء الاصطناعي الموالي، إلا إذا جرى توسيعها لتتضمَّن وضع تفضيلات البشر الآخرين في الاعتبار، إلى جانب تفضيلات المالك.

(٢-٢) الذكاء الاصطناعي النفعي

السبب وراء أنَّ لدينا فلسفة أخلاقية هو وجود أكثر من شخصٍ على كوكب الأرض. وعادة ما يُسمَّى النهج الأكثر ارتباطًا بفهم الكيفية التي يجب بها تصميم نظم الذكاء الاصطناعي بـ «العواقبية»؛ أي فكرة أن الاختيارات يجب الحكم عليها تبعًا للنتائج المتوقعة. أما النهجان الأساسيان الآخران، فهما «أخلاق الواجب» و«أخلاق الفضيلة»، اللذان يهتمَّان بشدة بالطابع الأخلاقي للأفعال والأفراد، على التوالي، بعيدًا عن نتائج الاختيارات.[3] في غياب أي دليل على وجود وعيٍ ذاتيٍّ لدى الآلات، أعتقد أنه ليس من الحكمة إنشاء آلات تتمتع بالفضيلة أو تختار أفعالًا تتوافق مع قواعد أخلاقية إن كانت التبعات غير مرغوبٍ فيها على نحوٍ كبير بالنسبة للبشرية. دعني أصوغ الأمر على نحوٍ آخر: إننا نُنشئ آلاتٍ لتحقيق نتائج، ويجب أن نفضل إنشاء آلاتٍ تُحقِّق نتائج نُريدها.

هذا لا يعني أن الفضائل والقيم الأخلاقية غير ذات صلة؛ إنها، بالنسبة إلى الشخص النفعي، مبرَّرة بالنظر إلى النتائج والتحقيق الأكثر عمليةً لتلك النتائج. تعرض جون ستيوارت ميل لتلك النقطة في عمله «النفعية»:

> الطرح القائل بأن السعادة هي الغاية والهدف من الأخلاق لا يعني أنه لا يجب وضع طريق للوصول إلى ذلك الهدف أو أن الناس الذين يسعون إليه لا يجب نُصحهم باتخاذ اتجاهٍ معين دون الآخر. ... لا أحد يُجادل أن فنَّ الملاحة لا يقوم على علم الفلك لأنَّ البحارة لا يمكنهم الانتظار لحساب التقويم البحري. ولأن البحارة مخلوقات عقلانية، فهم يذهبون إلى البحر بحسابات جاهزة؛ وكل المخلوقات العقلانية تخرج إلى بحر الحياة وعقولها لديها مفاهيم محدَّدة عن المسائل الشائعة المتعلِّقة بالصواب والخطأ، إلى جانب العديد من المسائل الأكثر صعوبة المتعلِّقة بالحكمة والطيش.

تتوافق تلك الرؤية على نحوٍ تام مع الفكرة التي ترى أن الآلة المتناهية التي تواجه التعقيد الهائل للعالم الواقعي قد تنتج نتائج أفضل بالالتزام بقواعد أخلاقية وتبنِّي أسلوبٍ قويمٍ بدلًا من محاولة تحديد مسار الفعل الأمثل من الصفر. باستخدام نفس الطريقة، يُحقِّق غالبًا برنامج الشطرنج الفوز باستخدام مجموعةٍ من تسلسلات الحركات الافتتاحية القياسية وخوارزميات إنهاء اللعب ودالة تقييم وليس بمحاولة الوصول إلى طريقةٍ للفوز دون إرشادات «أخلاقية». إن النهج العواقبي أيضًا يُعطي بعض الأهمية لتفضيلات هؤلاء الذين يُؤمنون بقوةٍ بالحفاظ على إحدى قواعد الواجب؛ لأنَّ الحزن الناتج عن عدم الالتزام بتلك القاعدة يعدُّ إحدى النتائج الحقيقية. ومع ذلك، إنها ليست نتيجة ذات أهمية لا مُتناهية.

العواقبية مبدأ صعب الاعتراض عليه — على الرغم من محاولة الكثيرين فعل ذلك! — لأنه من غير المنطقي الاعتراض على العواقبية على أساس أنها ستكون لها تبعات غير مرغوب فيها. فلا يُمكن أن يقول أحد: «لكن إن اتَّبعت النهج العواقبي في الحالة الفلانية، فإنَّ هذا الأمر الفظيع حقًّا سيحدث!» إن أيًّا من تلك الإخفاقات سيكون ببساطة دليلًا على أن النظرية قد أُسيء تطبيقُها.

على سبيل المثال، افترض أن هاريت تُريد تسلُّق جبل إيفرست. قد يخشى أحدهم أن روبي العواقبي ببساطة سيحملها إلى أعلى ويضعها على قمة هذا الجبل، نظرًا لأن تلك

هي النتيجة المرغوبة بالنسبة لها. على نحو شبه مؤكَّد، ستعترض هاريت على تلك الخطة، لأنها ستحرمها من التحدي؛ ومن ثمَّ من النشوة التي تنتج عن النجاح في إتمام مهمة صعبة بالاستعانة بجهودها الفردية. والآن، من الواضح أن روبي العواقبي المُصمَّم على نحوٍ جيد سيفهم أن النتائج تتضمن كل تجارب هاريت، وليس الغاية النهائية فقط. قد يرغب في أن يكون متاحًا في حالة وقوع أي مكروه وأن يتأكَّد من أنها مُدرَّبة جيدًا ومزوَّدة بكل التجهيزات اللازمة، لكنه أيضًا قد يكون عليه قبول حق هاريت في تعريض نفسها لخطر الموت.

إن كُنا نُخطِّط لإنشاء آلات عواقبية، فالسؤال الذي يطرح نفسه هو: كيف نُقيِّم العواقب التي تؤثِّر على العديد من الأشخاص؟ إحدى الإجابات المعقولة تتمثَّل في إعطاء أهميةٍ متساوية لتفضيلات الجميع؛ بعبارة أخرى، تعظيم مجموع منافع الجميع. عادة ما تنسب تلك الإجابة للفيلسوف البريطاني المُنتمي للقرن الثامن عشر جيرمي بنثام[4] وتلميذه جون ستيوارت ميل،[5] الذي قدم الطرح الفلسفي للنفعية. يُمكن إرجاع جذور الفكرة الأساسية إلى أعمال الفيلسوف اليوناني القديم إبيقور وهي تظهر صراحة في «موتسي»، وهو كتاب يضمُّ كتاباتٍ منسوبة إلى الفيلسوف الصيني الذي يحمل نفس الاسم. إن موتسي كان نشطًا في نهاية القرن الخامس قبل الميلاد وروَّج لفكرة «جياناي»، التي ترجمت بطرقٍ مُختلفة على أنها «الرعاية الشاملة» أو «الحب العالمي»، مُعتبرًا إياها السمة المميزة للأفعال الأخلاقية.

إن للنفعية سمعة سيئة إلى حدٍّ ما، جزئيًّا بسبب بعض سوء الفهم البسيط بشأن ما تتبناه. (بالتأكيد ما يزيد الأمر سوءًا أن تعني كلمة «نفعي» الآتي: «مُصمَّم كي يكون نافعًا أو عمليًّا بدلًا من أن يكون جذابًا».) النفعية عادة ما ينظر إليها على أنها تتعارض مع الحقوق الفردية لأنَّ الشخص النفعي من المُفترض ألا يتردَّد في نزع أعضاء أي شخصٍ إن كان سينقذ حياة خمسة أشخاص آخرين؛ بالطبع، مثل هذه الفكرة ستجعل الحياة غير آمنة على نحوٍ غير مقبول للجميع على الكوكب؛ لذا، الشخص النَّفعي لن يفكر حتى فيها. النفعية أيضًا مرتبطة على نحوٍ غير صحيح بتعظيم غير جذاب إلى حدٍّ ما للثروة الكلية ويُعتقد أنها لا تهتمُّ كثيرًا بالشعر أو المعاناة. في واقع الأمر، نُسخة بنثام منها ركزت على وجه الخصوص على السعادة البشرية، في حين أن ميل أكد بثقة على أن اللذات الذهنية لها قيمة أكبر بكثير من اللذات الجسدية. (من الأفضل أن تكون إنسانًا غير راضٍ عن خنزير راضٍ.) إن «النفعية المثالية» لجي إي مور ذهبت حتى لأبعد من هذا؛ لقد دعا إلى تعظيم الحالات الذهنية للقيمة الداخلية، المُمثَّلة في التأمُّل الجمالي للجمال.

أعتقد أنه لا تُوجَد مدعاة لقيام فلاسفة النفعية بتحديد المحتوى المثالي للمنفعة البشرية أو التفضيلات البشرية. (وأعتقد أن تلك المدعاة حتى تقل في حالة باحثي الذكاء الاصطناعي.) يستطيع البشر فعل ذلك لأنفسهم. أثار الاقتصادي جون هورشاني وجهة النظر هذه من خلال مبدئه المتمثِّل في «استقلالية التفضيلات»:[6]

> عند تحديد الصواب والخطأ بالنسبة إلى شخص مُعيَّن، المعيار النهائي يُمكن أن يكون فقط رغباته وتفضيلاته.

إن «نفعية التفضيلات» لدى هورشاني من ثمَّ مُتوافقة تقريبًا مع المبدأ الأول للذكاء الاصطناعي النافع، والذي ينصُّ على أن الغاية الوحيدة للآلة هي تحقيق التفضيلات البشرية. يجب بالطبع على باحثي الذَّكاء الاصطناعي ألا يكونوا جزءًا من محاولة تحديد الماهية التي «يجب» أن تكون عليها التفضيلات البشرية! وهورشاني، شأنه شأن بنثام، يرى تلك المبادئ باعتبارها وسيلةً لتوجيه القرارات «العامة»؛ فهو لا يتوقَّع أن يكون الناس غيريِّين جدًّا إلى هذه الدرجة. ولا يتوقع كذلك أن يكونوا عقلانيِّين على نحوٍ تام؛ على سبيل المثال، قد تكون لهم رغبات قصيرة الأجل تتعارض مع «تفضيلاتهم الأعمق». وأخيرًا، اقترح تجاهُل تفضيلات هؤلاء الذين، مثل هاريت السادية المذكورة آنفًا، يتمنَّون بشدة الإضرار بمصلحة الآخرين.

قدم هورشاني أيضًا دليلًا إلى حدٍّ ما على أنَّ القرارات الأخلاقية المثالية يجب أن تُعظِّم من المنفعة المتوسِّطة عبر أيِّ مجموعةٍ من البشر.[7] وقد افترض مُسلَّمات ضعيفة إلى حدٍّ ما مُماثلة لتلك التي تقوم عليها نظرية المنفعة بالنسبة إلى الأفراد. (المسلمة الأساسية الإضافية تتمثَّل في أنه إن يكن كل أعضاء المجموعة غير مُبالين تجاه نتيجتَين، فإن أي كيان يعمل بالنيابة عن المجموعة يجب أن يكون غير مبالٍ تجاه هاتَين النتيجتَين.) من هذه المُسلَّمات، أثبت ما صار معروفًا باسم «مبرهنة التجميع الاجتماعي»؛ أي الكيان الذي يعمل بالنيابة عن مجموعةٍ من الأفراد يجبُ أن يُعظِّم من مزيجٍ خطِّي موزون من المنافع الخاصة بهم. وحاجج كذلك بأنَّ الكيان «غير الإنساني» يجب أن يستخدم أوزانًا مُتساوية.

تتطلَّب المُبرهنة افتراضًا إضافيًّا مُهمًّا (وغير مذكور)، وهو أن كل الأفراد لديها نفس الاعتقادات الواقعية المسبقة عن العالم والطريقة التي سيتطوَّر بها. لكنَّ أيَّ أبٍ يعرف أن هذا حتى غير صحيح فيما يتعلَّق بأبنائه، فضلًا عن الأفراد الذين من خلفيات وثقافات

اجتماعية مختلفة. ومن ثم، ماذا سيحدث عند اختلاف الأفراد في اعتقاداتهم؟ سيحدث شيء غريب جدًّا:[8] الوزن المعطى لمنفعة كل فرد يجب أن يتغيَّر بمرور الوقت، بالتناسُب مع مدى تلاؤم الاعتقادات المُسبقة لهذا الفرد مع الواقع المتطوِّر.

إن تلك الصيغة التي تبدو غير عادلة إلى حدٍّ كبير مألوفة جدًّا لأيِّ أب. دعنا نفترض أن الروبوت روبي طُلب منه رعاية الطفلين، أليس وبوب. تريد أليس الذهاب إلى السينما وهي متأكدة من أن الطقس سيكون مُمطرًا اليوم؛ أما بوب، على الجانب الآخر، فيُريد الذهاب إلى الشاطئ وهو متأكِّد من أن الطقس سيكون مُشمسًا. يُمكن أن يقول روبي: «سنذهب إلى السينما»، مما سيجعل بوب غير سعيد، أو أن يقول: «سنذهب إلى الشاطئ»، مما سيجعل أليس تشعُر بالحزن أو يُمكنه القول: «إن كان الطقس مُمطرًا فسنذهب إلى السينما؛ أما إن كان مشمسًا فسنذهب إلى الشاطئ». إن تلك الخطة الأخيرة ستجعل كلًّا من أليس وبوب سعيدًا لأنَّ الاثنين يؤمنان بصحة ما يعتقدانه.

(٢-٣) التحديات التي تواجه النفعية

النفعية هي أحد الطروح التي نتجت عن بحث البشرية الطويل المدى عن دليل أخلاقي، وهي تُعدُّ، من بين العديد من مثل هذه الطروح، الأكثر تحديدًا على نحوٍ واضح؛ ومن ثم، الأكثر عرضةً لوجود ثغرات فيها. بدأ الفلاسفة يكتشفون تلك الثغرات منذ أكثر من مائة عام. على سبيل المثال، تخيَّل جي إي مور، الذي اعترض على تأكيد بنثام على تعظيم اللذة، «عالَمًا لا يُوجَد فيه تقريبًا شيء سوى اللذة؛ لا معرفة ولا حب ولا استمتاع بالجمال ولا سمات أخلاقية».[9] ستجدُ لهذا صدًى مُعاصرًا في إشارة ستيوارت أرمسترونج إلى أن الآلات الخارقة المكلَّفة بتعظيم اللذة قد «تدفن الجميع في توابيت أسمنتية على قطرات هيروين».[10] إليك مثالًا آخر: في عام ١٩٤٥، اقترح كارل بوبر الهدف الجدير بالاحترام والثناء الخاص بتقليل المعاناة البشرية،[11] ورأى أنه من غير الأخلاقي مُبادلة ألم شخصٍ بلذة آخر؛ وردَّ آر إن سمارت بأن هذا يُمكن تحقيقه على أفضل نحوٍ بجعل الجنس البشري ينقرض.[12] وفي هذه الأيام، فكرة أن الآلة قد تُنهي المعاناة البشرية بإنهاء وجودنا تُعَدُّ محور الجدل حول الخطر الوجودي للذكاء الاصطناعي.[13] هناك مثال ثالث يتمثل في تأكيد جي إي مور على «واقعية» مصدر السعادة، مُعدلًا بذلك التعريفات السابقة التي بدا أن بها ثغرة تسمح بتعظيم السعادة من خلال الخداع الذاتي. إن الأمثلة المعاصرة لهذه النقطة تتضمن فيلم «المصفوفة» (ذا ماتريكس) (الذي يتحوَّل فيه واقعُ اليوم إلى

وهمٍ أنتجتْه المحاكاة الحاسوبية)، والأبحاث الحديثة على مشكلة الخداع الذاتي في التعلم المعزز.[14]

تلك الأمثلة، وغيرها، تُقنعني بأن مجتمع الذكاء الاصطناعي يجب أن ينتبه بشدة إلى نقاط الهجوم والهجوم المضاد التي تُثار في النقاشات الفلسفية والاقتصادية الخاصة بالنفعية؛ لأنها ذات صلة على نحو مُباشر بالمهمة الحالية. يتعلَّق اثنان من أهم تلك النقاشات، من وجهة نظر تصميم نُظُم ذكاء اصطناعي تُفيد العديد من الأشخاص، بالمقارنات بين منافع الأفراد ومقارنات المنافع عبر أحجام مجموعات سكانية مختلفة. لقد بدأ هذان النقاشان منذ ١٥٠ عامًا أو يزيد، مما يؤدي بالمرء للشك في أن انتهاءهما على نحوٍ مُرضٍ قد لا يكون سهلًا على الإطلاق.

إن النقاش بشأن المقارنات بين منافع الأفراد مُهم لأن روبي لا يُمكنه تعظيم مجموع منفعتي أليس وبوب إلا إذا كان بالإمكان جمع هاتَين المنفعتَين؛ ويُمكن فعل هذا فقط إن كانتا قابلتين للقياس على نفس المقياس. حاجج عالم المنطق والاقتصاد البريطاني المُنتمي إلى القرن التاسع عشر ويليام ستانلي جيفنز (الذي يُعدُّ أيضًا مخترع كمبيوتر ميكانيكي مبكر يُسمَّى البيانو المنطقي) في عام ١٨٧١ بأن تلك المقارنات مستحيلة:[15]

> إن قابلية تأثُّر أحد العقول، بحسب علمنا، قد تكون أكبر ألف مرة من تلك الخاصة بعقل آخر. لكن، نظرًا إلى أن تلك القابلية كانت مختلفة بنسبة مُتشابهة في كل الاتجاهات، فيجب ألا يكون بإمكاننا أبدًا اكتشاف الاختلاف الأعمق. من ثمَّ، كلُّ عقل يكون مُستغلقًا بالنسبة إلى العقول الأخرى، وإيجاد قاسم مشترك فيما يتعلق بالشعور غير ممكن.

كان الاقتصادي الأمريكي كينيث أرو، الذي يُعدُّ مؤسس نظرية الاختيار الاجتماعي الحديثة والحائز على جائزة نوبل في عام ١٩٧٢، صارمًا على نحوٍ مُماثل:

> إن وجهة النظر المتَّخذة هنا هي أن المقارنة بين منافع الأفراد لا معنى لها، وفي الحقيقة، لا معنى مُتعلِّقًا بمقارنات الرفاهية في قابلية المنفعة الفردية للقياس.

الصعوبة التي يُشير إليها جيفنز وأرو تتمثَّل في عدم وجود طريقة واضحة لتحديد ما إذا كانت أليس تقدر وخزات الدبابيس والمصاصات بالقيمتين −١ و+١ أم −١٠٠٠ و+١٠٠٠ في ضوء تجربتها الذاتية للسعادة. في كلتا الحالتين، ستدفع من أجل الحصول

على مصاصة لتجنُّب الوخز بالدبوس. في واقع الأمر، إن كانت أليس آليًّا شبيهًا بالإنسان، فإن سلوكها الخارجي قد لا يختلف رغم عدم وجود تجربة ذاتية للسعادة على الإطلاق.

في عام ١٩٧٤، أشار الفيلسوف الأمريكي روبرت نوزيك إلى أنه حتى إن كان بالإمكان عمل مقارنات بين منافع الأفراد، فإن تعظيم مجموع المنافع سيظلُّ يُعدُّ فكرةً سيئة لأنه سيصطدم بما يسمى بـ «وحش المنفعة»؛ وهو الشخص الذي تكون تجارب اللذة والألم لديه أكثر قوة عدة مرات من تلك الخاصة بالأشخاص العاديِّين.[16] مثل هذا الشخص يُمكن أن يؤكد أن أي وحدة إضافية من الموارد ستنتج زيادة أكبر في المجموع الكُلي للسعادة البشرية إن أُعطيت له بدلًا من الآخرين؛ في واقع الأمر، «أخذ» موارد من الآخرين لصالح وحش المنفعة سيكون فكرة جيدة أيضًا.

قد تبدو هذه نتيجة غير مرغوب فيها على نحوٍ واضح، لكن العواقبية في حدِّ ذاتها لا يمكن أن تفيد هنا: المشكلة تكمن في كيفية قياس مرغوبية النتائج. أحد الردود الممكنة تتمثل في أن وحش المنفعة مجرد شيء نظري؛ إذ لا يُوجَد أشخاص مثل هذا. لكن هذا الرد على الأرجح لن يفيد أيضًا؛ فبنحوٍ ما، «كل» البشر وحوشُ منفعةٍ مُقارنةً، لنقُل، بالجرذان والبكتيريا، وهذا هو السبب وراء عدم اهتمامنا الكبير بتفضيلات الجرذان والبكتيريا عند وضع السياسة العامة.

إذا كانت فكرة أن الكيانات المُختلفة لديها مقاييس منفعة مختلفة مضمنة بالفعل في طريقة تفكيرنا، فيبدو من الممكن تمامًا أن يكون لدى الأشخاص المختلفين مقاييس مختلفة أيضًا.

هناك رد آخر يتمثَّل في ندب الحظ والعمل على أساس الافتراض الذي يرى أن الجميع لديهم المقياس نفسه، حتى لو لم يكونُوا كذلك.[17] كما يُمكن للمرء محاولة استكشاف الأمر من خلال الوسائل العلمية التي لم تكن متاحةً لجيفنز، مثل قياس مستويات الدوبامين أو درجة الإثارة الكهربية للعصبونات المرتبطة باللذة والألم، والسعادة والبؤس. إذا كانت الاستجابات الكيميائية والعصبية لأليس وبوب فيما يتعلَّق بالمصاصات مُتطابقة إلى حدٍّ كبير، وكذلك استجاباتهم السلوكية (الابتسام وأصوات لعق الشفاه وغير ذلك)، فيبدو من الغريب الإصرار مع ذلك على أن درجتي استمتاعهما الشخصية تختلفان بعامل قدره ألف أو مليون. وأخيرًا، يستطيع المرء استخدام الأشياء المشتركة الشائعة مثل الوقت (التي لدينا جميعًا منها، تقريبًا، نفس القدر)؛ على سبيل المثال، بمقارنة المصاصات ووخزات الدبابيس، لنقل، بفترة انتظار إضافية قدرها ٥ دقائق في صالة المغادرة الخاصة بالمطار.

أنا أقلُّ تشاؤمًا بكثيرٍ من جيفنز وأرو. إنني أعتقد أنه من المُفيد بالفعل المقارنة بين منافع الأفراد وأرى أن المقاييس قد تختلف ولكن ليس بالأساس بعوامل كبيرة للغاية وأنَّ الآلات يُمكن أن تبدأ باعتقادات مُسبقة عامة على نحو معقول فيما يتعلَّق بمقاييس التفضيلات البشرية وتتعلَّم المزيد عن مقاييس الأفراد بالملاحظة بمرور الوقت، ربما بربط الملاحظات الطبيعية بنتائج أبحاث علم الأعصاب.

النقاش الثاني — المُتعلِّق بمقارنات المنفعة عبر المجموعات السكانية ذات الأحجام المختلفة — يكون مُهمًّا عندما يكون للقرارات تأثير على من سيُوجد في المستقبل. على سبيل المثال، في فيلم «المنتقمون: الحرب الأزلية» (أفنجرز: إنفنيتي وور)، شخصية ثانوس تطور وتنفذ النظرية التي تقول إنه لو قلَّ عددُ سكان العالم بمقدار النصف، فسيكون البشر الباقون أكثر سعادةً بمقدارٍ يزيد عن الضعف. وهذه هي نوعية الحسابات الساذجة التي أعطت مذهب النفعية سمعةً سيئة.[18]

نفس المسألة — فيما عدا الأحجار الأزلية والميزانية الجبارة — نُوقشت في عام ١٨٧٤ على يد الفيلسوف البريطاني هنري سيدجويك في عمله الشهير «أساليب الأخلاق».[19] خلص سيدجويك، في اتفاق واضح مع ثانوس، إلى أن الاختيار الصحيح كان تعديل حجم السكان حتى يجري الوصول إلى السعادة الإجمالية القصوى. (من الواضح أن هذا لا يعني زيادة عدد السكان دون حدود؛ لأن الجميع في نقطة معيَّنة سيتضوَّرُون جوعًا حتى الموت؛ ومن ثمَّ سيكونون في غاية التعاسة.) في عام ١٩٨٤، تناول الفيلسوف البريطاني ديريك بارفيت تلك المسألة ثانية في عمله الرائد «الأسباب والأشخاص».[20] حاجج بارفيت بأنه بالنسبة إلى أي وضع يكون فيه عدد الأشخاص السعداء للغاية بالمجموعة السكانية ن، فهناك (طبقًا للمبادئ النفعية) وضع أفضل فيه يكون عدد الأشخاص الأقل سعادة بقليل ٢ن. يبدو هذا معقولًا جدًّا. لسوء الحظ، هناك أيضًا ما يُسمى بمنحدر زلق. فبتكرار العملية، نصل إلى ما يُطلق عليه «الاستنتاج البغيض» (وهو مُصطلح له جذور تعود إلى العصر الفيكتوري): ويعني أن الوضع الأكثر مرغوبية هو ذلك الذي يُوجَد فيه سكان كُثر، والذين لجميعهم حياة بالكاد تستحق العيش.

كما يُمكن أن تتخيَّل، إن هذا الاستنتاج مثير للجدل. بارفيت نفسه صارع لأكثر من ثلاثين عامًا لإيجاد حلٍّ لمُعضلته، لكن دون أن ينجح في ذلك. أعتقد أنه ينقصنا بعض المُسلمات الجوهرية، المناظرة لتلك الخاصة بالتفضيلات العقلانية على نحوٍ فردي، للتعامل مع التفضيلات عبر المجموعات السكانية ذات الأحجام ومُستويات السعادة المختلفة.[21]

من المهم أن نحلَّ هذه المشكلة؛ لأنَّ الآلات التي لديها تبصُّر كافٍ قد تكون قادرةً على التفكير في مسارات فعلٍ تؤدِّي إلى أحجام مجموعات سكانية مختلفة، تمامًا كما فعلت الحكومة الصينية من خلال سياسة الطفل الواحد التي أقرتها في عام ١٩٧٩. من المُحتمل للغاية، على سبيل المثال، أن نطلب من نظم الذكاء الاصطناعي المساعدة في وضع حلول لمشكلة تغير المناخ العالمي، وقد تتضمَّن تلك الحلول وضع سياساتٍ تميل إلى الحد من النمو السكاني أو حتى تقليله.[22] على الجانب الآخر، إن قرَّرنا أن المجموعات السكانية الأكبر حقًّا أفضل وأعطينا أهمية كبيرة لرخاء المجموعات السكانية البشرية التي ربما تكون كبيرة؛ وذلك على مدى قرونٍ من الآن، فسنحتاج للعمل على نحو أكبر من أجل إيجاد طرق لتجاوز حدود كوكبنا. وإن أدَّت حسابات الآلات إلى الاستنتاج البغيض أو نقيضه — عدد سكان قليل أفرادُه سعداء على نحوٍ مثاليٍّ — فسيكون علينا الشعور بالندم لعدم تحقيقنا التقدم المنشود في هذه المشكلة.

حاجج بعض الفلاسفة بأننا قد نحتاج لاتخاذ قرارات في ظل حالة من عدم اليقين الأخلاقي؛ أي عدم اليقين بشأن النظرية الأخلاقية الملائمة التي ستُستخدم في اتخاذ القرارات.[23] أحد الحلول يتمثَّل في تخصيص بعض الاحتمال لكل نظرية أخلاقية واتخاذ القرارات باستخدام «قيمة أخلاقية متوقَّعة». لكن ليس من الواضح إن كان من المعقول تخصيص احتمالات للنظريات الأخلاقية بنفس الطريقة التي نُطبِّق بها الاحتمالات على طقس الغد. (ففي النهاية، ما احتمال أن تكون وجهة نظر ثانوس صحيحة تمامًا؟) وحتى إن كان هذا معقولًا، فالاختلافات التي ربما تكون كبيرة بين توصيات النظريات الأخلاقية المتنافسة تعني أن إنهاء عدم اليقين الأخلاقي — تحديد النظرية الأخلاقية التي تتجنَّب التبعات غير المقبولة — يجب أن يحدث «قبل» اتخاذ تلك القرارات المهمَّة أو العهد بذلك للآلات.

دعنا نكُن متفائلين ونفترض أن هاريت في النهاية ستحلُّ تلك المُشكلة وغيرها من المشكلات الناشئة عن وجود أكثر من شخصٍ على كوكب الأرض. جرى تنزيل خوارزميات مُناسبة تقوم على الغيرية والمساواة في الروبوتات عبر جميع أنحاء العالم. وهناك مظاهر احتفال وموسيقى سعيدة في كل مكان. وبعد ذلك، تعود هاريت إلى المنزل:

روبي: عود حميد! أكان يومًا طويلًا؟

هاريت: نعم، لقد كان العمل شاقًّا حقًّا، ولم تتسنَّ لي حتى فرصة تناول الغداء.

روبي: إذن، لا بدَّ أنك جائعة بشدة!

هاريت: أكادُ أموت جوعًا! هل بإمكانك إعداد عشاء لي؟

روبي: هناك شيء أنا بحاجة لإخبارك به ...

هاريت: ما هو؟ لا تقلُ لي إن الثلاجة خاوية!

روبي: لا، هناك أناس في الصومال في حاجة عاجلة أكثر للمُساعدة. أنا سأُغادر الآن. رجاء أعدِّي عشاءك بنفسك.

في حين أن هاريت ربما تكون في غاية الفخر بروبي وبإسهاماتها في سبيل جعله آلة محترمة ومتميزة، فقد لا تستطيع منع نفسها من التساؤل عن السبب وراء دفعها مبلغًا كبيرًا في شراء روبوت أوَّل تصرُّف مُهمٍّ له هو تركها. فعليًّا، بالطبع، لن يشتري أحد مثل هذا الآلي؛ ومن ثم لن يجري تصنيع مثل هذه الروبوتات، ولن تكون هناك أي منفعة للبشرية منه. دعنا نُطلق على هذا «المشكلة الصومالية». فمن أجل أن ينجح نظام آلي نفعي بالكامل، يجب أن نجد حلًّا لتلك المشكلة. سيحتاج روبي لأن يكون لديه قدر من الولاء لهاريت على الخصوص؛ ربما، قدر مُرتبط بالمبلغ الذي دفعتْه هاريت من أجل شرائه. في كل الأحوال، إن أراد المجتمع أن يساعد روبي أناسًا آخرين بجانب هاريت، فعليه أن يُعوِّضها فيما يتعلق بحقها في الاستفادة من جهود روبي. ومن المحتمل إلى حدٍّ كبيرٍ أن يُوجَد تنسيق بين الروبوتات بحيث لا ينزل الجميع إلى الصُّومال في وقتٍ واحد؛ وفي هذه الحالة، قد لا يحتاج روبي في النهاية للذهاب إلى هناك. أو ربما تظهر بعض الأنواع الجديدة تمامًا من العلاقات الاقتصادية للتعامل مع وجود مليارات الكيانات الغيرية على نحوٍ تامٍّ في العالم (وهو الأمر الذي بالتأكيد يُعدُّ غير مسبوق).

(٣) حسدُ البشر وشرُّهم ومراعاتهم للغير

تتجاوز التفضيلات البشرية اللذَّة والبيتزا. إنها بالتأكيد تمتدُّ إلى مصلحة الآخرين. حتى آدم سميث، الذي يعدُّ أبا الاقتصاد الذي عادة ما يُقتبس كلامه عندما تكون هناك حاجة لتبرير الأنانية، بدأ كتابه الأول بالتأكيد على الأهمية القصوى للاهتمام بشئون الآخرين:[24]

> مهما كانت درجة الأنانية المُفترضة في الإنسان، فمن الواضح أن هناك بعض المبادئ في طبيعته تجعلُه يهتمُّ بمصلحة الآخرين وتجعل سعادتهم ضرورية له، رغم أنه لا يستمدُّ أي شيء من ذلك سوى مُتعة رُؤيتها. ومن ذلك الشعور بالشفقة أو التعاطف، تلك العاطفة التي نشعر بها تجاه بؤس الآخرين عندما

نراه أو يجعلُنا الآخرون نتصوَّرها على نحوٍ واضح للغاية. إن كوننا غالبًا ما نستمدُّ الحزن من أحزان الآخرين لهو أمر واقع واضح للغاية لا يتطلَّب أي أمثلة لإثباته.

في اللغة الاقتصادية الحديثة، الاهتمام بالآخرين عادة ما يندرج تحت موضوع «الغيرية».[25] إن نظرية الغيرية مُصاغة على نحوٍ جيِّد إلى حدٍّ ما، ولها تبعات مُهمة على سياسة الضرائب من ضمن أمور أخرى. ويجب القول إن بعض الاقتصاديِّين يتعاملون مع الغيرية باعتبارها شكلًا آخر من الأنانية المصمَّمة لتمدَّ المُعطي «بتوهُّج دافئ».[26] هذا بالطبع احتمال تحتاج الروبوتات لأن تكون على وعيٍ به عند تفسيرها للسلوك البشري، ولكن دعنا في هذا المقام نُؤمن بغيريَّة البشر ونفترض أنهم يهتمُّون بالآخرين بالفعل.

أسهل طريقة للنظر في الغيرية هي تقسيم تفضيلات الفرد إلى نوعَين؛ هما: التفضيلات المتعلقة بمصلحته الشخصية والتفضيلات المتعلقة بمصلحة الآخرين. (هناك جدل كبير حول ما إذا كان بالإمكان الفصل بين الاثنين على نحوٍ تامٍّ، لكنَّني سأُنحِّي هذا الأمر جانبًا.) تشير المصلحة الشخصية إلى سمات حياة الفرد الذاتية، مثل المأوى والشعور بالدفء والحصول على الطعام والأمان وما إلى ذلك، المرغوب فيها في حدِّ ذاتها وليس بالنظر إلى سمات حياة الآخرين.

لجعل هذا المفهوم واضحًا أكثر، دعنا نفترض أن العالم يعيش به شخصان هما أليس وبوب. تتألف منفعة أليس الإجمالية من قيمة مصلحتها الشخصية مضافًا إليها عامل ما وهو هـبأ مع ضرب الناتج في قيمة مصلحة بوب الشخصية. إن «عامل الاهتمام» هـبأ يشير إلى مدى اهتمام أليس بمصلحة بوب. وعلى نحوٍ مماثل، تتألف منفعة بوب الإجمالية من قيمة مصلحته الشخصية مضافًا إليها عامل اهتمام ما وهو هـأب مع ضرب الناتج في قيمة مصلحة أليس الشخصية، بحيث يشير هـأب إلى مدى اهتمام بوب بمصلحة أليس.[27] يُحاول روبي مساعدة كل من أليس وبوب، مما يعني (دعنا نقُل) تعظيم مجموع المنفعتين. ومن ثم يحتاج روبي إلى الانتباه ليس فقط إلى المصلحة الفردية لكلٍّ منهما، ولكن أيضًا إلى كيف يهتمُّ كلٌّ منهما بمصلحة الآخر.[28]

إن علامة معاملي الاهتمام هـبأ وهـأب مهمة جدًّا. على سبيل المثال، إذا كان هـبأ موجبًا، فإن أليس «مراعية للغير»؛ أي تستمد بعض السعادة من تحقق مصلحة بوب. وكلما كان هذا العامل موجبًا أكثر، كانت أليس على استعداد للتضحية ببعض مصلحتها الشخصية

لمساعدة بوب. وإن كان هذا العامل صفرًا، فأليس أنانية تمامًا؛ أي إن كان بإمكانها النجاة من العقاب، ستُحوِّل أي قدر من الموارد من بوب وتُوجِّهه إليها، حتى وإن تركت بوب في حالة بائسة ويتضوَّر جوعًا. عندما يجد روبي النفعي أن أليس أنانية وأن بوب مُراعٍ للغير، من الواضح أنه سيحمي بوب من أسوأ أفعال أليس. من المثير للاهتمام أن التوزان النهائي في الغالب سيخدم مصلحة أليس أكثر من مصلحة بوب، لكن قد تكون لديه سعادة إجمالية أكبر لأنه يهتم بمصلحتها. قد تشعر بأن قرارات روبي غير عادلة على نحو كبير إن خدمت مصلحة أليس أكثر من مصلحة بوب فقط لأنه أكثر مراعاة للغير منها: هل عليه أن يستاء من هذه النتيجة ويصبح غير سعيد؟[29] حسنًا، قد يفعل ذلك لكن هذا سيكون نموذجًا مختلفًا؛ وهو النموذج الذي يتضمَّن مُصطلحًا للاستياء فيما يتعلَّق بالاختلافات في خدمة المصلحة. في نموذجنا البسيط، سيتقبل بوب النتيجة. في واقع الأمر، في حالة التوازن، سيُقاوم بوب أيَّ محاولة لتحويل الموارد من أليس إلى نفسه، نظرًا لأن هذا سيقلل من سعادته الإجمالية. إن ظننت أنَّ هذا غير واقعي بالمرة، فتأمَّل الحالة التي تكون فيها أليس ابنة بوب الحديثة الولادة.

الحالة المُلغزة حقًّا بالنسبة إلى روبي ستكون عندما يكون العامل $هـ_{بأ}$ سالبًا؛ ففي هذه الحالة، تكون أليس شريرة حقًّا. سأستخدم هنا المصطلح «الغيرية السالبة» للإشارة إلى مثل هذه التفضيلات. وكما هو الحال مع هاريت السادية المذكورة قبل ذلك، هذا لا يتعلَّق بالأنانية والحقد الشائعَين، بحيث تكون أليس سعيدةً لاقتناص نصيب بوب من الكعكة حتى تزيد نصيبها. الغيرية السالبة تعني أن أليس تستمدُّ سعادتها من عدم تحقُّق مصلحة الآخرين، حتى وإن بقيت مصلحتها الشخصية كما هي دون تغيير.

في البحث الذي قدم فيه هورشاني مصطلح نفعية التفضيلات، نسب الغيرية السالبة إلى «السادية والحسد والاستياء والحقد» وحاجج بأنها يجب تجاهلها عند حساب المجموع الإجمالي للمنفعة البشرية في أي مجموعة سكانية:

> لا يمكن لأي قدْر من الاهتمام بشخصٍ ما أن يفرض عليَّ التزامًا أخلاقيًّا بمساعدته في إيذاء شخصٍ آخر.

يبدو هذا أحد المجالات الذي من المعقول فيه بالنسبة لُمصمِّمي الآلات الذكية أن يُحاولوا (بحذَر) التأثير على النتائج من أجل تحقيق العدالة.

لسوء الحظ، الغيرية السالبة أكثر شيوعًا بكثيرٍ مما قد يتوقَّعه المرء. إنها لا تنشأ من السادية والحقد[30] بقدر ما تنشأ من الحسد والاستياء وعاطفتهما العكسية، والتي أُسمِّيها «الفخر» (وذلك لعدم وجود كلمة أدق يُمكنني استخدامها). إذا كان بوب يحسد أليس، فهو يشعر بالحزن بسبب «الاختلاف» فيما بينهما فيما يتعلَّق بتحقُّق مصلحتهما؛ فكلما كان الاختلاف أكبر، زاد حزنه. على الجانب الآخر، إن كانت أليس فخورة بتفوُّقها على بوب، فإنها تستمد السعادة ليس فقط من تحقُّق مصلحتها الشخصية، وإنما أيضًا من حقيقة أن مصلحتها تحققت على نحوٍ أكبر من مصلحته. ومن السهل إثبات، على نحوٍ رياضي، أن الفخر والحسد يعملان بنفس الطريقة تقريبًا مثل السادية؛ فهما يجعلان أليس وبوب يستمدان السعادة فقط من عدم تحقُّق مصلحة كلٍّ منهما لأن عدم تحقُّق مصلحة بوب يزيد من فخر أليس، في حين أن عدم تحقق مصلحة أليس يُقلِّل من حسد بوب.[31]

ذكر لي جيفري ساكس، عالم اقتصاد التنمية المعروف، قصةً أوضحت تأثير هذه الأنواع من التفضيلات في تفكير الناس. كان ساكس في بنجلاديش بعد فترةٍ وجيزة من تعرُّض إحدى مناطق البلاد لفيضان كبير. كان يتحدَّث إلى أحد المزارعين الذي فقد منزله وحقوله وكل حيواناته وأحد أبنائه. وقال له: «أنا حزين بشدة من أجلك؛ لا بدَّ أنك تعيس للغاية.» كان رد المزارع: «لا، على الإطلاق.» وأضاف: «أنا سعيد جدًّا لأنَّ جاري الملعون فقد زوجته وكل أبنائه أيضًا.»

التحليل الاقتصادي للفخر والحسد — بخاصة في سياق المكانة الاجتماعية والاستهلاك التفاخري — برز من خلال عمل عالم الاجتماع الأمريكي ثورشتاين فيبلن الذي عرض عمله «نظرية الطبقة المترفة» الذي ظهر في عام ١٨٩٩ التبعات السيئة لهذه التوجهات.[32] وفي عام ١٩٧٧، نشر عالم الاقتصاد البريطاني فريد هيريش كتابه «الحدود الاجتماعية للنمو»[33] الذي قدم فيه فكرة «السِّلع الموضعية». إن السلعة الموضعية هي أي شيء — والذي قد يكون سيارة أو منزلًا أو ميدالية أوليمبية أو نوعَ تعليم أو دخلًا أو لكنة — يستمد قيمته المدركة ليس فقط من مزاياه الجوهرية ولكن أيضًا من خصائصه النسبية، بما في ذلك خصائص الندرة والتفوق على الآخرين. إن السعي وراء السلع الموضعية، الذي يقوده الفخر والحسد، يكون له طابع لعبة المجموع الصفري، بمعنى أن أليس لا يُمكنها تحسين موضعها النِّسبي دون تأزيم الموضع النسبي لبوب، والعكس صحيح. (لا يبدو أن هذا يمنع إنفاق مبالغ ضخمة في هذا المسعى.) يبدو أن السِّلع الموضعية

كثيرة في الحياة الحديثة، لذا ستحتاج الآلات إلى فهم أهميتها الكلية في تفضيلات الأفراد. علاوة على ذلك، يرى مُنظِّرو نظرية الهوية الاجتماعية أن العضوية في جماعة والانتماء إليها والمكانة الإجمالية للجماعة بالنسبة إلى الجماعات الأخرى تُعدُّ عناصر أساسية لتقدير البشر لذواتهم.[34] ومن ثم من الصعب فهم السلوك البشري دون فهم كيف يرى الأفراد أنفسهم كأعضاء في جماعات، سواء كانت تلك الجماعات أنواعًا بيولوجية أو أُممًا أو جماعاتٍ عرقية أو أحزابًا سياسية أو مهنًا أو أسرًا أو مُشجِّعين لفريق كرة قدم مُعين.

كما هو الحال مع السادية والحقد، قد نرى أن روبي يجبُ ألا يعطي أهميةً كبيرة للفخر والحسد أو لا يُعيطهما أهمية على الإطلاق في خططه لمساعدة أليس وبوب. مع ذلك، هناك بعض الصعوبات في هذا الطرح. فنظرًا لأنَّ الفخر والحسد يتعارضان مع اهتمام أليس بمصلحة بوب، فقد لا يكون من السهل الفصل بينهما. ربما تكون أليس مهتمَّة بشدة بمصلحة بوب، لكنها تحسده أيضًا؛ فمن الصعب تمييز أليس هذه من أليس أخرى لديها اهتمام قليل بمصلحة بوب، ولكن ليس لديها حسد على الإطلاق تجاهه. بالإضافة إلى ذلك، في ضوء شيوع الفخر والحسد في التفضيلات البشرية، من المُهم التفكير بحذَر شديد في تبعات تجاهلهما. فقد يكونان ضروريَّين لتقدير الذات، خاصة في شكليهما الإيجابيَّين؛ احترام الذات وتقدير الآخرين.

دعني أعيد التأكيد على نقطة ذكرتها قبل ذلك، وهي أن الآلات المصممة على نحوٍ ملائم «لن تتصرف مثل من تُلاحظهم»، حتى وإن كانت تلك الآلات تتعلَّم تفضيلات شياطين ساديين. من المُمكن، في واقع الأمر، أننا نحن البشر إن وجدنا أنفسنا في الوضع غير المألوف المُتمثل في التعامل مع كيانات غيرية بالكامل على نحوٍ يوميٍّ، قد نتعلم أن نكون أناسًا أفضل؛ أي نكون أكثر غيرية ويقلُّ توجيه الفخر والحسد إلى أفعالنا.

(٤) غباء البشر وعاطفيتهم

ليس المقصود من عنوان هذا القسم الإشارة إلى مجموعة فرعية معينة من البشر. إنه يشير إلينا جميعًا. إننا جميعًا أغبياء على نحوٍ غير معقول في ضوء المعيار المُتعذِّر الوصول إليه، الخاص بالعقلانية التامة، وكلنا مُعرَّضون لتقلبات العواطف المُختلفة التي، إلى حدٍّ كبير، تتحكم في سلوكنا.

دعنا نبدأ بالغباء. يُعظِّم الكيان العقلاني تمامًا من التحقيق المتوقَّع لتفضيلاته عبر كل الحيوات المستقبلية المُمكنة التي يُمكن أن يختار أن يعيشها. لا يُمكنني كتابة عددٍ يصف تعقُّد مشكلة اتخاذ القرار هذه، لكنني أجد التجربة الفكرية التالية مفيدةً في هذا الشأن. أولًا: لاحظ أن عدد اختيارات التحكم الحركي التي يتَّخذها أيُّ شخصٍ في حياته تصلُ إلى عشرين تريليون. (انظر المُلحق «أ» للاطلاع على الحسابات التفصيلية.) ثانيًا: دعنا نرى إلى أي مدى ستُوصِّلُنا القوة المفرطة بمساعدة كمبيوتر سيث لويد المحمول الذي يُلامس أقصى حدود القدرات الفيزيائية المُمكنة، الذي هو أسرع مليار تريليون تريليون مرةً من أسرع كمبيوتر في العالم. سنعهد إليه بمهمة عدِّ كلِّ التسلسلات الممكنة للكلمات الإنجليزية (ربما كتدريبٍ إحمائي لمكتبة بابل التي يُصورها خورخ لويس بورخس)، وسنجعلُه يعمل لمدة عام. السؤال الآن: ما طول التسلسلات التي يمكنه عدُّها في ذلك الوقت؟ ألف صفحة من النصوص؟ مليون صفحة؟ لا. ١١ كلمة فقط. يعطيك هذا لمحة عن صعوبة تصميم أفضل حياة مُمكنة بها عشرون تريليون فعل. باختصار، إننا بعيدون جدًّا عن العقلانية تمامًا مثل بُعد البزاق عن السيطرة على المركبة الفضائية «إنتربرايز» التي تسير بسرعة ٢٥٠ مليون كيلومتر في الثانية. نحن «ليس لدَينا على الإطلاق أي فكرة» عن الشكل الذي ستكون عليه الحياة المختارة على نحوٍ عقلاني.

يدلُّ هذا على أن البشر سيتصرَّفون عادةً بطرُق تتعارض مع تفضيلاتهم الشخصية. على سبيل المثال، عندما خسر لي سيدول مباراته في لعبة جو أمام برنامج «ألفا جو»، لعب حركة واحدة أو أكثر «أكَّدت» أنه سيخسر، واستطاع البرنامج (في بعض الحالات على الأقل) اكتشاف قيامه بذلك. لكن سيكون من الخطأ أن يستنتج البرنامج أن ليو سيدول يُفضِّل الخسارة. بدلًا من ذلك، سيكون من المعقول استنتاج أن ليو سيدول يُفضِّل الفوز لكن لديه بعض القصور الحوسبي الذي منعه من اختيار الحركة الصحيحة في كل الحالات. ومن ثم، من أجل فهم سلوك ليو سيدول واكتساب معلوماتٍ عن تفضيلاته، يجب على الروبوت الذي يتبع المبدأ الثالث (مصدر المعلومات الأساسي للتفضيلات البشرية هو السلوك البشري) معرفة بعض المعلومات عن العمليات المعرفية التي تُنتج هذا السلوك. فهو لا يستطيع افتراض أن سيدول عقلاني.

هذا يُمثِّل مشكلةً بحثية مهمة جدًّا بالنسبة إلى باحثي الذكاء الاصطناعي وعلم النفس وعلم الأعصاب؛ وهي فهم ما يكفي عن المعرفة البشرية[35] بحيث يُمكننا (أو بالأحرى، يمكن لآلاتنا النافعة) القيام «بالهندسة العكسية» للسلوك البشري للوصول إلى التفضيلات

الأساسية العميقة، بالمدى الذي هي عليه. استطاع البشر القيام بقدْر من هذا، حيث عرفوا قيمهم من الآخرين من خلال بعض المساعدة من علم البيولوجيا، لذا، يبدو هذا مُمكنًا. إن لدى البشر ميزة؛ بإمكانهم استخدام بنيتهم المعرفية لمحاكاة تلك الخاصة بغيرهم من البشر دون معرفة ماهية تلك البنية؛ «إن أردتَ شيئًا ما، فسأفعل نفس ما تفعلُه أمي تمامًا، لذا، لا بدَّ أن أُمي تُريد هذا الشيء».

ليس لدى الآلات تلك الميزة. إن بإمكانها مُحاكاة الآلات الأخرى بسهولة، ولكن ليس البشر. ومن غير المُحتمل أن يكون لديها قريبًا وصولًا لنموذجٍ كامل للمعرفة البشرية، سواء عام أو مُصمَّم لأفراد بعينهم. بدلًا من ذلك، من الأفضل من الناحية العملية النظر إلى الطرق الأساسية التي ينحرف بها البشر عن العقلانية ودراسة كيفية تعلم التفضيلات من السلوك الذي يبدي تلك الانحرافات.

هناك اختلاف واحد واضح بين البشر والكيانات العقلانية والذي يتمثَّل في أننا، في أي لحظة، لا نختار من بين كل الخطوات الأولى المُمكنة لكل الحيوات المستقبلية الممكنة. ونحن حتى لسنا قريبين من هذا. بدلًا من هذا، نحن في العادة غارقون في تسلسُل متداخل بشدة من «الروتينات الفرعية». بوجه عام، نحن نسعى إلى تحقيق أهداف قريبة الأجل بدلًا من تعظيم تحقيق التفضيلات عبر حيوات مُستقبلية، ويُمكننا التصرُّف فقط تبعًا لحدود الروتين الفرعي الموجودين فيه في الوقت الحاضر. أنا الآن، على سبيل المثال، أكتب هذه الجملة: يُمكنني اختيار كيفية الاستمرار بعد علامة النقطتَين، لكن لم يخطُر لي أبدًا أن أتساءل إن كان عليَّ التوقُّف عن كتابة الجملة والانضمام إلى أحد البرامج التدريبية الخاصَّة بغناء الراب على الإنترنت أو إضرام النار في المنزل والاتِّصال بشركة التأمين أو فعل أيِّ شيءٍ من ملايين الأشياء التي «يُمكنني» فعلها بعد ذلك. إن الكثير من تلك الأشياء الأخرى قد تكون بالفعل أفضل مما أفعلُه، لكن، في ضوء تسلسل الالتزامات الخاص بي، يبدو الأمر وكأن تلك الأشياء الأخرى غير موجودة.

إذن، يبدو أن فهم الفعل البشريِّ يتطلَّب فهم تسلسل الروتينات الفرعية هذا (الذي قد يكون فرديًّا إلى حدٍّ كبير): الروتين الفرعي الذي يُنفِّذه الشخص حاليًّا، والهدف القريب الأجل الذي يجري السعي من أجل تحقيقه داخل الروتين الفرعي هذا، وكيفية ارتباطهما بالتفضيلات الطويلة الأجل الأكثر عُمقًا. بوجهٍ عام أكثر، يبدو أن تعلم التفضيلات البشرية يتطلب معرفة الهيكل الفعلي للحيوات البشرية. ما هي كل الأشياء التي يمكن أن نقوم بها نحن البشر، سواء على نحوٍ فردي أو مُشترك؟ ما الأنشطة المميزة للثقافات وأنواع الأفراد

المختلفة؟ إن هذين السؤالين مُثيران للاهتمام ويحتاجان إلى البحث. من الواضح أنهما ليس لهما إجابة ثابتة لأننا نحن البشر نضيف أنشطة وهياكل سُلوكية جديدة لُمخزوننا منهما طوال الوقت. لكن حتى الإجابات الجزئية والمؤقتة ستكون مفيدةً جدًّا لكل أنواع النظم الذكية المصممة لمساعدة البشر في حيواتهم اليومية.

هناك خاصية واضحة أخرى للأفعال البشرية والتي تتمثَّل في أنها عادة ما تقودها العاطفة. في بعض الحالات، هذا شيء جيد؛ فالعواطف مثل الحب والعرفان بالجميل تعدُّ بالطبع جزئيًّا جزءًا أساسيًّا من تفضيلاتنا، والأفعال التي تنتج عنها يمكن أن تكون عقلانية، حتى وإن لم تكن مقصودة على نحوٍ تام. وفي حالات أخرى، تؤدي الاستجابات العاطفية إلى أفعالٍ حتى نحن البشر الأغبياء نرى أنها ليست عقلانية على الإطلاق؛ بعد حدوثها، بالطبع. على سبيل المثال، إن هاريت الغاضبة والمحبطة التي ضربت أليس العنيدة البالغة من العمر عشرة أعوام قد تندم على ما قامت به على الفور. يجب على روبي، الملاحظ لفعل هاريت، (كما هو مُتوقع وإن لم يكن في كل الأحوال) أن يعزو هذا التصرف إلى الغضب والإحباط وعدم ضبط النفس وليس إلى السادية المقصودة لذاتها. وحتى يتم ذلك، يجب أن يكون لدى روبي بعض الفهم للحالات العاطفية البشرية، بما في ذلك أسبابها وكيفية تطورها عبر الوقت استجابة للمُثيرات الخارجية وتأثيراتها على الفعل. بدأ علماء الأعصاب يضعون أيديهم على آليات بعض الحالات العاطفية وعلاقاتها بالعمليات المعرفية الأخرى،[36] وهناك بعض الأبحاث المُفيدة عن الطرق الحاسوبية المتعلِّقة باكتشاف الحالات العاطفية البشرية وتوقعها والتعامل معها،[37] لكن ما زال هناك الكثير الذي يجبُ معرفته. مرةً أخرى، الآلات لديها مشكلة فيما يتعلَّق بالعواطف؛ فهي لا يُمكنها إنتاج محاكاةٍ داخلية لأي تجربةٍ لتحديد الحالة العاطفية التي سيُنتجها.

بالإضافة إلى تأثير العواطف على أفعالنا؛ فهي تكشف معلومات مفيدة عن تفضيلاتنا الأساسية. على سبيل المثال، ربما كانت أليس الصغيرة ترفُض أداء فروضها المنزلية، وهاريت غاضبة ومحبطة لأنها تريد حقًّا أن يكون لأليس أداء جيد في المدرسة وأن تكون لديها فرصة أفضل في الحياة مما توفَّرت لهاريت. إذا كان روبي مُستعدًّا لفهم هذا — حتى إن لم يختبر ذلك بنفسه — فقد يتعلَّم الكثير من أفعال هاريت غير العقلانية. لذا، يجب أن يكون من المُمكن إنشاء نماذج أولية للحالات العاطفية البشرية تكفي لتجنُّب الأخطاء الأكثر شناعة في استنتاج التفضيلات البشرية من السلوك.

(٥) هل للبشر تفضيلات حقًّا؟

إن الافتراض الأساسي الذي يقوم عليه هذا الكتاب يتمثَّل في وجود حيوات مُستقبلية نسعى إلى الوصول إليها، وأخرى نرغب في تجنُّبها، مثل التعرض للانقراض على المدى القصير أو التحوُّل إلى مزارع بطاريات بشرية على غرار ما حدث في فيلم «المصفوفة». بهذا المعنى، نعم، بالتأكيد البشر لديهم تفضيلات. لكن بمجرَّد أن نغوص في تفاصيل الكيفية التي سيفضلون أن تكون عليها حيواتهم، تصبح الأمور أكثر غموضًا.

(٥-١) عدم اليقين والخطأ

تتمثَّل إحدى الخصاص الواضحة للبشر، إن لاحظتها، في أنهم دائمًا لا يعرفون ما يريدون. على سبيل المثال، يكون للأشخاص المُختلفين استجابات مختلفة تجاه فاكهة الدوريان؛ البعض يجد «أنها تتجاوز كل أنواع الفاكهة الأخرى في العالم في الطعم»[38] في حين أن آخرين يُشبِّهونها بـ «ماء الصرف الصحي والقيء الجاف والرائحة الكريهة التي يُخرجها الظربان والماسحات الجراحية المُستعملة».[39] تجنَّبتُ مُتعمِّدًا تجربة فاكهة الدوريان قبل كتابة هذا الكتاب، حتى أستطيع الحفاظ على حياديتي في هذه النقطة: أنا ببساطة لا أعرف إلى أي الفريقَين سأنتمي. نفس الشيء يمكن أن يُقال بالنسبة إلى العديد من الأشخاص الذين يُفكِّرون في حيواتهم العملية المستقبلية أو شركاء حياتهم المستقبليين أو أنشطة ما بعد التقاعُد المستقبلية وهكذا.

هناك على الأقل نوعان من عدم اليقين فيما يتعلَّق بالتفضيلات. الأول عدم يقين معرفي حقيقي، مثل ذلك الذي خابرته فيما يتعلَّق بتفضيلي لفاكهة الدوريان.[40] لن يُنهي أي قدْر من التفكير هذا النوع من عدم اليقين. هناك حقيقة تجريبية للأمر، ويمكنني معرفة المزيد من خلال تجربة بعض حبات تلك الفاكهة أو مقارنة الحمض النووي الخاص بي مع ذلك الخاص بمُحبي تلك الفاكهة وكارهيها أو غير ذلك. وينشأ النوع الثاني عن بعض القصور الحوسبي: عند النظر لوضعين في لعبة جو، أنا غير متأكِّد أيهما أُفضل لأن تبعات كل منهما خارج نطاق قدرتي على التحديد تمامًا.

ينشأ عدم اليقين أيضًا من حقيقة أن الاختيارات التي تُتاح أمامنا عادة ما تكون محدَّدة على نحوٍ غير كامل؛ أحيانًا على نحوٍ غير كامل تمامًا بحيث يُمكن اعتبارها بالكاد كاختيارات. على سبيل المثال، عندما تُصبح أليس على وشك إنهاء دراستها الثانوية، قد

يعرض عليها أحد مُستشاري التوظيف الاختيار ما بين أن تعمل في وظيفة «أمينة مكتبة» أو «عاملة بمنجم فحم»؛ قد تقول، على نحو معقول تمامًا: «أنا لستُ على يقين فيما يتعلَّق بتفضيلي في هذا الشأن». هنا، عدم اليقين ينشأ من عدم يقين معرفي خاصٍّ بتفضيلاتها فيما يتعلق، لنقُل، بغبار الفحم في مقابل غبار الكتب؛ ومن عدم يقين حوسبي وهي تُحاول جاهدة تحديد كيف قد تنجح في كل من هذين الاختيارين المتعلِّقين بعملها؛ ومن عدم يقين عادي فيما يتعلق بالعالم، مثل شُكوكها بشأن الصلاحية الطويلة الأجل لمنجم الفحم المحلي الخاص بها.

لتلك الأسباب، إنها لفكرة سيئة أن نربط التفضيلات البشرية باختيارات بسيطة بين خيارات موصوفة على نحوٍ غير كامل من المتعذِّر تقييمها وتتضمن عناصر من المرغوبية غير المعلومة. توفر تلك الاختيارات مُؤشِّرًا غير مباشر على التفضيلات المتضمنة، لكنها ليست جزءًا من تلك التفضيلات. وهذا ما جعلني أستعرض مفهوم التفضيلات فيما يتعلَّق «بالحيوات المستقبلية»؛ على سبيل المثال، بتخيُّل أن بإمكانك مشاهدة، على نحو مضغوط، فيلمَين مُختلفين لحياتك المُستقبلية ثمَّ التعبير عن أيهما تُفضِّل (ارجع إلى الفصل الثاني). إن التجربة الفكرية هذه بالطبع من المستحيل تنفيذها على أرض الواقع، لكن يُمكن للمرء أن يتصوَّر أنه في العديد من الحالات سينشأ تفضيل واضح قبل فترة طويلة من معرفة كل تفاصيل كلِّ فيلم ومُشاهدتها بالكامل. قد لا تعرف مقدمًا أيهما ستُفضِّل، حتى لو أُعطيت مُلخَّصًا لحبكة كلٍّ منهما؛ لكن هناك إجابة للسؤال الفعلي، بناءً على ما أنت عليه الآن، تمامًا كما أن هناك إجابة على سؤال ما إذا كنت ستُحب فاكهة الدوريان عندما تُجرِّبها.

إن حقيقة أنك قد تكون غير مُتيقِّن بشأن تفضيلاتك الشخصية لا تُسبب أي مشكلات بعينها فيما يتعلَّق بالطرح المُعتمد على التفضيلات الخاص بالذكاء الاصطناعي النافع على نحوٍ مُثبت. في الحقيقة، هناك بالفعل بعض الخوارزميات التي تضع في اعتبارها عدم يقين روبي وهاريت بشأن تفضيلات هاريت وتسمح باحتمالية أنَّ هاريت ربما تكتسب معلومات بشأن تفضيلاتها في نفس الوقت الذي يفعل فيه روبي ذلك.[41] وكما أن عدم يقين روبي بشأن تفضيلات هاريت يُمكن تقليله بملاحظة سلوك هاريت، فإن عدم يقين هاريت بشأن تفضيلاتها الشخصية يُمكن تقليله بملاحظة ردود أفعالها تجاه التجارب. لا يجب أن يكون هذان النوعان من عدم اليقين مُرتبطَين على نحوٍ مباشر؛ كما أن روبي ليس بالضرورة أقل تيقنًا من هاريت فيما يتعلق بتفضيلاتها. على سبيل المثال، قد يكون

روبي قادرًا على اكتشاف أن هاريت لدَيها استعداد وراثي مُسبق قوي للاشمئزاز من رائحة فاكهة الدوريان. في هذه الحالة، سيكون لديه عدم يقين قليل للغاية بشأن تفضيلها لتلك الفاكهة، حتى لو ظلَّت على جهلٍ تامٍّ بهذا الأمر.

إن كانت هاريت «غير مُتيقنة» بشأن تفضيلاتها الخاصة بالأحداث المستقبلية، فمن المرجح إلى حدٍّ كبير أن تكون أيضًا «مخطئة». على سبيل المثال، قد تكون مُقتنعةً بأنها لن تُحبَّ الدوريان (أو، لنقل، البيض الأخضر أو لحم فخذ الخنزير)؛ ومن ثمَّ ستتجنَّبها مهما حدث، لكنها قد تجد في النهاية أنها رائعة — إن وضع أحد عن طريق الخطأ البعض منها في سلطة الفاكهة الخاصة بها في أحد الأيام. ومن ثمَّ لا يستطيع روبي افتراض أن أفعال هاريت تعكس معرفة دقيقة بتفضيلاتها الشخصية؛ فالبعض قد يكون مُعتمدًا تمامًا على التجربة، في حين أن البعض الآخر قد يكون قائمًا على نحوٍ رئيسيٍّ على الافتراض أو الانحياز أو الخوف من المجهول أو التعميمات التي ليست لها أُسُس قوية.[42] إن روبي اللبق على نحوٍ ملائم يُمكن أن يكون مفيدًا للغاية لهاريت فيما يتعلَّق بتنبيهها لمثل هذه المواقف.

(٥-٢) التجربة والذكريات

بعض علماء النفس شكَّك في صحة فكرة أن هناك ذاتًا تفضيلاتها مُهيمنة بالطريقة التي اقترحها مبدأ استقلالية التفضيلات الخاصُّ بهورشاني. من أبرز علماء النفس هؤلاء زميلي السابق في بيركلي دانيال كانمان. يُعدُّ كانمان، الذي حصل على جائزة نوبل لعام ٢٠٠٢ لعمله في مجال الاقتصاد السلوكي، واحدًا من أكثر المفكرين تأثيرًا في موضوع التفضيلات البشرية. وكتابه الذي ظهر حديثًا «التفكير، السريع والبطيء»[43] يعرض ببعض التفصيل سلسلة من التجارب التي أقنعته بوجود ذاتين — «الذات المُستشعرة» و«الذات المُتذكِّرة» — تتعارض تفضيلاتهما.

الذات المُستشعرة هي تلك التي قيست باستخدام «مقياس اللذة»، الذي تخيَّل الاقتصاديُّ البريطاني المُنتمي إلى القرن التاسع عشر فرانسيس إدجوورث أنه «أداة كاملة على نحوٍ مثاليٍّ، آلة نفسية فيزيائية، تُسجِّل باستمرار ذروة اللذة التي يختبرها الفرد، على نحوٍ دقيق وفقًا لحكم الوعي».[44] وفقًا للنفعية القائمة على اللذَّة، القيمة الإجمالية لأيِّ تجربة بالنسبة لأيِّ فرد هي ببساطة مجموع القيم القائمة على اللذة لكلِّ لحظة

أثناء التجربة. وينطبق هذا المفهوم، بقدْرٍ مُتساوٍ، على تناول الأيس كريم أو عيش حياة بأكملها.

إن الذات المُتذكِّرة، على الجانب الآخر، هي تلك التي تتولَّى القيادة عند اتخاذ أي قرار. تختار تلك الذات تجارب جديدة اعتمادًا على «ذكريات» تجارب سابقة ومرغوبيَّتها. تقترح تجارب كانمان أن الذات المُتذكِّرة لديها أفكار مُختلفة جدًّا عن الذات المُستشعرة.

تتضمَّن أبسط تلك التجارب في فهمها غمْر يد أحد المبحُوثين في الماء البارد. هناك نظامان مختلفان؛ في الأول، يكون الغمرُ لمدة ٦٠ ثانية في ماءٍ درجة حرارته ١٤ درجة مئوية؛ وفي الثاني، يكون لمدة ٦٠ ثانية في ماء درجة حرارته ١٤ درجة مئوية ثم لمدة ٣٠ ثانية في ماء درجة حرارته ١٥ درجة مئوية. (درجات الحرارة هذه مُماثلة لدرجات حرارة المحيط في شمال كاليفورنيا؛ وهي باردة بالقدر الكافي لارتداء الجميع تقريبًا لبذلة غوص في الماء.) قال كلُّ المبحوثين إن التجربة كانت غير سارة. وبعد تجربة كلا النظامَين (أيًّا كان الترتيب، مع وجود ٧ دقائق فيما بينهما)، طُلب من المبحوث اختيار أيهما سيودُّ تكراره. فضَّل الغالبية العظمى من المبحوثين تكرار النظام الثاني بدلًا من النظام الأول.

افترض كانمان أن النظام الثاني، من وجهة نظر الذات المُستشعرة، لا بد أنه «بالطبع أسوأ» من النظام الأول؛ لأنَّه يتضمَّن النظام الأول، «إلى جانب تجربة غير سارة أخرى». ومع ذلك، اختارته الذات المتذكرة. ربما تسأل عن السبب.

يتمثَّل تفسير كانمان في أن الذات المتذكرة تنظُر إلى الأمر، من خلال نظارة ملونة على نحو غريب بعض الشيء، مهتمة بنحو أساسي بقيمة «الذروة» (أعلى أو أقل قيمة للذة) وقيمة «النهاية» (قيمة اللذَّة في نهاية التجربة). يجري في الغالب تجاهُل مدة الأجزاء المختلفة للتجربة. إن مُستوى عدم الراحة الخاص بالذروة لكلٍّ من النظامَين متساوٍ، لكن مُستوى النهاية مختلف: في حالة النظام الثاني، الماء أكثر دفئًا بمقدار درجة واحدة. إن قيَّمَتِ الذاتُ المُتذكِّرة التجاربَ من خلال قيمتَي الذروة والنهاية، بدلًا من جمع قيم اللذة عبر الوقت، فإنَّ النظام الثاني سيكون أفضل، وهذا ما جرى التوصُّل إليه. يبدو أن نموذج الذروة والنهاية يُفسِّر العديد من النتائج الأخرى الغريبة على نحوٍ مُتساوٍ في الأدبيات الخاصة بالتفضيلات.

يبدو أن كانمان (ربما على نحوٍ ملائم) مُتحيِّر فيما يتعلق بالنتائج التي توصَّل إليها. إنه يؤكد على أن الذات المتذكِّرة «قد ارتكبت ببساطة خطأً»، واختارت التجربة الخاطئة لأن ذاكرتها معيبة وغير كاملة؛ إنه يرى هذا باعتباره «خبرًا سيئًا للمؤمنين

بعقلانية الاختيار». على الجانب الآخر، كتب يقول: «لا يُمكن دعم أي نظرية عن الرفاهية تتجاهل ما يُريده الناس». افترض، على سبيل المثال، أن هاريت قد جرَّبت نوعَي المشروبات الغازية الشهيرين وأنها تُفضل الآن بقوة أحدهما؛ سيكون من الغريب إجبارها على تناول النوع الآخر اعتمادًا على جمع قراءات مقياس لذةٍ ما مأخوذة في كل تجربة.

حقيقة الأمر أنه لا يُوجَد قانون «يتطلَّب» تعريف تفضيلاتنا فيما يتعلق بالتجارب من خلال مجموع قِيَم اللذة عبر الوقت. صحيح أن النماذج الرياضية القياسية تُركِّز على تعظيم مجموع المكافآت،[45] لكن الدافع الأصلي وراء هذا كان الملاءمة الرياضية. جاءت التبريرات لاحقًا في شكل افتراضاتٍ فنية ترى أنه من العقلانية اتخاذ القرار بناءً على جمع المكافآت،[46] لكن تلك الافتراضات الفنية لا يجب أن تكون صحيحة في الواقع. افترض، على سبيل المثال، أن هاريت تختار بين نتيجتَين لقيم اللذَّة، هما: [١٠، ١٠، ١٠، ١٠، ١٠] و[٠، ٠، ٤٠، ٠، ٠]. من المُمكن تمامًا أن تُفضِّل التسلسل الثاني؛ فلا يُوجد قانون رياضي يُمكن أن يُجبرها على اتخاذ اختياراتٍ اعتمادًا على المجموع بدلًا من، لنقُل، القيمة القُصوى.

يعترف كانمان أن الوضع يتعقَّد أكثر بسبب الدَّور المحوري للتوقُّع والذكريات في الرفاهية. إن ذكرى تجربة سارَّة واحدة — يوم زواج المرء أو ميلاد طفل أو عصر يوم قُضي في قطف التوت الأسود وصُنع المربَّى — يُمكن أن تدعم المرء في سنوات العمل الشاق والإحباط. إن الذات المتذكرة ربما تُقيم ليس فقط التجربة في حدِّ ذاتها، وإنما أيضًا تأثيرها الإجمالي على القيمة المستقبلية للحياة من خلال تأثيرها على الذكريات المستقبلية. وعلى الأرجح إن الذات المتذكرة وليس المُستشعرة هي أفضل حكم على ما سيجري تذكره.

(٥-٣) الزمن والتغيير

غني عن البيان أن الأشخاص الراشدين في القرن الحادي والعشرين لن يرغبُوا في تقليد تفضيلات، لنقُل، المجتمع الروماني في القرن الثاني، الحافل بالقتل بسبب المُصارعة البشرية من أجل التسلية العامة، والاقتصاد القائم على العبودية والمجازر الوحشية للشعوب المهزومة. (لا حاجة لنا باستعراض الأمور الواضحة المقابلة لتلك السمات في المجتمع المعاصر.) تتطوَّر مقاييس الأخلاق بوضوح بمرور الوقت مع تطور حضارتنا أو انحدارها، إن شئت القول. هذا يُشير، بدوره، إلى أن الأجيال المستقبلية قد تستهجن توجُّهاتنا الحالية، لنقُل، تجاه التعامُل مع الحيوانات. لهذا السبب، من المُهم أن تكون

الآلات المكلَّفة بتنفيذ التفضيلات البشرية قادرة على الاستجابة للتغييرات التي تحدُث في تلك التفضيلات بمرور الوقت بدلًا من الاستمرار على نفس التفضيلات. إن المبادئ الثلاثة المعروضة في الفصل السابع تستوعب تلك التغييرات بطريقة طبيعية، لأنها تتطلب أن تتعلَّم وتُنفِّذ الآلات تفضيلات البشر الحاليين — الكثير منهم، الذين كلهم مختلفون — بدلًا من مجموعة واحدة مثالية من التفضيلات أو تفضيلات مُصمِّمي الآلات الذين ربما يكونون قد ماتوا منذ فترة طويلة.[47]

إن احتمالية حدوث تغييرات في التفضيلات الأساسية للمجموعات السكانية البشرية عبر الزمن بطبيعة الحال تلفت الانتباه إلى المسألة المتعلِّقة بالطريقة التي تتكوَّن بها تفضيلات كل فرد ومرونة تفضيلات البالغين. إن تفضيلاتنا بالتأكيد تتأثر بجوانبنا البيولوجية: على سبيل المثال، إننا في الغالب نتجنَّب الألم والجوع والعطش. لكن جوانبنا البيولوجية ظلت ثابتة إلى حدٍّ ما، لذا، التفضيلات المُتبقية يجب أن تكون قد نشأت عن مؤثِّرات ثقافية وعائلية. من المُحتمل جدًّا أن الأطفال يُنفِّذُون باستمرار نوعًا من التعلم المعزَّز العكسي للتعرف على تفضيلات الآباء والأقران حتى يُفسِّروا سلوكهم، وبعد ذلك يتبنَّى الأطفال تلك التفضيلات وتُصبح خاصَّةً بهم. وحتى كبالغين، تتطوَّر تفضيلاتُنا بسبب تأثير الإعلام والحكومة والأصدقاء وأرباب الأعمال وتجاربنا الشخصية المُباشرة. قد يكون صحيحًا، على سبيل المثال، أنَّ الكثير من مُؤيِّدي ألمانيا النازية لم يبدءوا مسيرتهم كساديِّين متعطشين للإبادة الجماعية ونقاء العرق.

يُمثِّل تغيير التفضيلات تحديًا لنظريات العقلانية على المُستوى الفردي والمجتمعي. على سبيل المثال، يبدو أن مبدأ هورشاني الخاص باستقلالية التفضيلات يقول إن الجميع له الحق في امتلاك التفضيلات التي يُريدها ولا يحقُّ لأي شخص آخر أن يُغيرها. مع ذلك، وبعيدًا عن كون التفضيلات قابلة للتغيير، فإنها يجري تغييرها وتعديلها طوال الوقت، من خلال كل تجربةٍ يمر بها المرء. لا يسعُ الآلات إلا تعديل التفضيلات البشرية لأنَّ الآلات تُعدِّل التجارب البشرية.

من المهم، على الرغم من كونه أحيانًا صعبًا، التَّفرقة بين تغيير التفضيلات وتحديث التفضيلات، وهو الأمر الذي يحدُث عندما تتعلَّم هاريت غير المُتيقنة في البداية المزيد عن تفضيلاتها الشخصية من خلال التَّجربة. يُمكن أن يملأ تحديث التفضيلات الفجوات في المعرفة الذاتية وربما يؤكد أكثر التفضيلات التي كانت في السابق مؤقَّتة وذات أساسٍ ضعيف. إن تغيير التفضيلات، على الجانب الآخر، ليس عمليةً تنتُج عن امتلاك أدلةٍ

إضافية عن التفضيلات الفعلية للمرء. في الحالة القصوى، يُمكنك تخيل أنه ناتج عن تناول المُخدِّرات أو حتى الخضوع لجراحة دماغية؛ فهو ينشأ عن عمليات قد لا نفهمُها أو حتى نُوافق عليها.

يعدُّ تغيير التفضيلات مُشكلةً لسببَين على الأقل. السبب الأول هو أنه ليس من الواضح التفضيلات التي يجب أن تُهيمن عند اتخاذ أحد القرارات: التفضيلات التي تكون لدى هاريت في وقت اتخاذ القرار أم تلك التي ستكون لديها أثناء وبعد الأحداث التي تنتُج عن قرارها. في مجال علم الأخلاق البيولوجية، على سبيل المثال، تُعَدُّ هذه مُعضلةً واقعيةً جدًّا لأنَّ تفضيلات الناس بشأن التدخُّلات الطبية والرعاية في مرحلة الاحتضار تتغيَّر، عادة على نحوٍ هائل، بعد أن يُصبحُوا مرضى بشدَّة.[48] وبافتراض أن تلك التغييرات لم تنتج بسبب ضعف القُدرات العقلية، فتفضيلات من هي التي يجب احترامها؟[49]

السبب الثاني لكون تغيير التفضيلات مُشكلةً هو أنه يبدو أنه ليس هناك أساس عقلاني واضح لتغيير المرء لتفضيلاته (مقارنةً بتحديثها). إن كانت هاريت تُفضِّل شيئًا عن شيء آخر، لكن قد تختار المرور بتجربة تعرف أنها سينتج عنها تفضيل الشيء الثاني على الأول، فلماذا يجب من الأساس أن تفعل ذلك؟ سيكون الناتج هو أنها ستختار حينها الشيء الثاني، الذي لا تُريده حاليًّا.

إنَّ مسألة تغيير التفضيلات تظهر على نحوٍ دراميٍّ في أسطورة أوليس وحوريات البحر. إن حوريات البحر مخلوقات خيالية غناؤها يُغوي البحَّارة ويجعل مصيرهم الموت على صخور جزر معينة في البحر المتوسِّط. أمر أوليس، الذي كان يرغب في الاستماع إلى غناء الحوريات، بحارته بسدِّ آذانه بالشمع وربطه بصارية السفينة، وطلب منهم عدم إطاعة توسُّلاته اللاحقة بفكِّه تحت أيِّ ظرف. من الواضح أنه كان يُريد من البحارة احترام التفضيلات التي كانت لديه في البداية، وليس تلك التي ستكون لديه بعد إغواء الحُوريات له. تلك الأسطورة أصبحت عنوان كتاب للفيلسوف النرويجي جون إلستر،[50] الذي يتناول ضعف الإرادة والتحديات الأخرى للفكرة النظرية الخاصَّة بالعقلانية.

لماذا قد تسعى أيُّ آلة ذكية عن قصدٍ لتعديل تفضيلات البشر؟ الإجابة بسيطة جدًّا، وهي: لجعل التفضيلات أسهل في تحقيقها. لقد رأينا هذا في الفصل الأول في حالة تحسين معدَّل النقر في وسائل التواصُل الاجتماعي. أحد الردود قد تتمثَّل في القول بأنَّ الآلات يجبُ أن تتعامل مع التفضيلات البشرية باعتبارها شيئًا مُقدَّسًا؛ لا يُمكن أن يسمح لأيِّ شيءٍ بتغيير التفضيلات البشرية. لسوء الحظ، هذا مُستحيل تمامًا. إن وجود روبوت مُساعد مُفيد من المُحتمَل أن يكون له تأثير على التفضيلات البشرية.

يتمثَّل أحد الحُلول المُمكنة في تعلُّم الآلات «للتفضيلات التعريفية» البشرية؛ أي التفضيلات الخاصة بأنواع عمليات تغيير التفضيلات التي قد تكون مقبولةً أو غير مقبولة. لاحظ هنا استخدام «عمليات تغيير التفضيلات» بدلًا من «تغييرات التفضيلات». يرجع هذا إلى أنَّ الرغبة في تغيير الفرد لتفضيلاته في اتِّجاهٍ مُعيَّن عادةً ما يكون مُساويًا لامتلاك هذا التفضيل بالفعل؛ الشيء المطلوب بالفعل في تلك الحالة هو القُدرة على «تنفيذ» التفضيل على نحوٍ أفضل. على سبيل المثال، إن قالت هاريت: «أُريد لتفضيلاتي أن تتغيَّر بحيث لا أفضل الكعك كما أفعل الآن»، فلديها بالفعل تفضيل لُمستقبل تستهلك فيه كعكًا أقل؛ ما تُريده حقًّا هو تغيير بِنيتها المعرفية بحيث يعكس سلوكها على نحوٍ أكبر هذا التفضيل.

أقصد بـ «التفضيلات الخاصة بأنواع عمليات تغيير التفضيلات التي قد تكون مقبولةً أو غير مقبولة»، على سبيل المثال، وجهة النظر التي قد تؤدِّي بالمرء للوصول إلى تفضيلات «أفضل» من خلال السفر حول العالم والتعرُّف على مجموعة متنوعة من الثقافات أو المشاركة في أنشطة جماعة فكرية نابضة بالحياة تستكشف على نحوٍ تامٍّ نطاقًا كبيرًا من التقاليد الأخلاقية أو تخصيص بعض الوقت للتأمُّل والتفكير العميق في الحياة ومعناها. سأُطلق على تلك العمليات «التفضيلات الحيادية»، بمعنى أن المرء لا يتوقَّع أن العملية ستُغيِّر تفضيلاته في أيِّ اتجاهٍ معين، مع إدراك أن بعضها قد يتعارض بشدَّة مع هذا التوصيف.

بالطبع، ليس كل عمليات التفضيلات الحيادية مرغُوبة؛ على سبيل المثال، يتوقَّع القليل من الناس تطوير تفضيلات «أفضل» من خلال ضرب أنفسهم على رءوسهم. إن تعريض الذات لعملية تغيير تفضيلات مقبولة يُناظر تنفيذ تجربة لمعرفة القليل عن كيف يعمل العالم؛ أنت لن تعرف أبدًا مُقدمًا النتيجة التي ستئول إليها التجربة، ولكنَّك تتوقَّع، مع ذلك، أن تكون في وضع أفضل في حالتك الذهنية الجديدة.

يبدو أن فكرة أن هناك سُبلًا مقبولة لتعديل التفضيلات ترتبط بفكرة أن هناك طرقًا مقبولة لتعديل السُّلوك والتي بمُقتضاها، على سبيل المثال، رب العمل سيضبط موقف الاختيار بحيث يتَّخذ الناس اختيارات «أفضل» فيما يتعلَّق بالادِّخار من أجل التقاعُد. عادة ما يُمكن القيام بهذا بالتعامل مع العوامل «غير العقلانية» التي تؤثر على الاختيار، بدلًا من تقييد الاختيارات أو العقاب على الاختيارات «السيئة». عرض كتاب «الوكزة» الذي وضعه الاقتصاديُّ ريتشارد ثالر والباحث القانوني كاس صانشتاين، لنطاق عريض من

الطُّرُق والفُرص التي من المفترض أنها مقبولة والتي يُمكنها «التأثير على سلوك الناس حتى تجعل حياتهم أطول وأكثر صحة وأفضل».

من غير الواضح ما إذا كانت طرق تعديل السلوك تُعدل حقًّا السلوك فقط. إن استمرَّ، بعد اختفاء الوكزة، السُّلوك المعدَّل، وهو الأمر الذي من المُفترض أن يُعدَّ الناتج المرغوب فيه لمثل هذه التدخُّلات — فقد تغيَّر شيء في البنية المعرفية للفرد (الشيء الذي يُحوِّل التفضيلات المعنيَّة إلى سلوك) أو في التفضيلات المعنية للفرد. ومن المُحتمل جدًّا أن الشيء المُتغيِّر يكون مزيجًا من الاثنين. لكن الأمر الواضح هو أن استراتيجية الوكزة تفترض أنَّ الجميع يشاركون تفضيلًا خاصًّا بالحياة «الأطول والأكثر صحة والأفضل»؛ كل وكزة قائمة على تعريف محدَّد للحياة «الأفضل»، والذي يبدو أنه يتعارض مع السِّمة الأساسية لاستقلالية التفضيلات. قد يكون من الأفضل، بدلًا من ذلك، تصميم عمليات تفضيلات حيادية مُعاونة تُساعد الناس على جعل قراراتهم وبنياتهم المعرفية مُتناسقة على نحو أفضل مع تفضيلاتهم المعنية. على سبيل المثال، من المُمكن تصميم عمليات معاونة معرفية تركز على التَّبعات ذات المدى الأطول للقرارات وتعلم الناس كيفية إدراك جذور تلك التبعات في الحاضر.[51]

إنَّ الحاجة إلى الوصول إلى فهمٍ أفضل للعمليات التي بمُقتضاها تتكوَّن وتتشكَّل التفضيلات البشرية تبدو واضحة لعدةَّ أسباب؛ أهمها أن مثل هذا الفهم سيُساعد في تصميم آلاتٍ تتجنَّب التغييرات العرضية وغير المرغوب فيها في التفضيلات البشرية من النوع الذي تقوم به خوارزميات انتقاء المُحتوى على مواقع التواصل الاجتماعي. عندما نُصبح مزوَّدين بمثل هذا الفهم، فإننا بالتأكيد سنسعى إلى إحداث تغييرات ستُؤدِّي إلى عالمٍ «أفضل».

قد يُحاجج البعض بأننا يجب أن نُوفِّر فُرصًا أكبر بكثير لتجارب «تحسين» التفضيلات الحيادية؛ مثل السفر والجدال والتدريب في مجال التفكير النَّقدي والتحليلي. قد نُوفِّر، على سبيل المثال، فرصًا لكلِّ طالب ثانوي للعيش لبضعة أشهُر في ثقافتين أُخريين — على الأقل — مُختلفتَين عن ثقافته.

لكنَّنا على نحوٍ شبه مؤكَّد سنرغب في المُضيِّ قُدُمًا أبعد من ذلك؛ على سبيل المثال، بإجراء إصلاحات اجتماعية وتعليمية تزيد من معامل الغيرية — أي الوزن الذي يُعطيه كل فرد لمصلحة الآخرين — مع تقليل معاملات السادية والفخر والحسد. هل سيكون هذا هدفًا جيدًا؟ هل سنستعينُ بآلاتنا لمساعدتنا في تنفيذ هذه العملية؟ إن الأمر مُغرٍ

بالتأكيد. في واقع الأمر، كتب أرسطو نفسُه يقول: «المسعى الأساسي للسياسة هو تكوين المواطنين لشخصية معيَّنة وجعلهم صالحين وميَّالين للقيام بأفعالٍ نبيلة». دعنا نقُل فقط إن تلك هي المخاطر المرتبطة بهندسة التفضيلات المقصودة على نطاقٍ واسع. يجبُ أن نسير متَّخذين الحيطة القُصوى.

الفصل العاشر

هل حُلَّت المشكلة؟

إن نجحنا في بناء نُظم ذكاءٍ اصطناعيٍّ نافعة على نحوٍ مُثبت، فسنُقلِّل خطر احتمالية فقداننا للتحكُّم في الآلات الخارقة. يُمكن للبشرية حينها أن تستمرَّ في تطورها وتجني الفوائد التي تكاد تكون غير مُتخيَّلة التي ستنشأ من القُدرة على السيطرة على ذكاء أكبر بكثير في قيادة حضارتنا لمزيد من التقدُّم. سنتحرَّر من قرونٍ من العبودية كُنا فيها مثل روبوتات تعمل في مجال الزراعة والصناعة والعمل الإداري، وسيكون بإمكاننا استغلال الفُرَص التي تُوفِّرها لنا الحياة على النحو الأمثل. وفي ضوء ذلك العصر الذهبي، سننظُر إلى حياتنا في الوقت الحاضر تمامًا كما تخيَّل توماس هوبز الحياة بدون حكومة: منعزلة وفقيرة وشرِّيرة وبهيمية وقصيرة.

أو ربما لا يكون هذا هو الحال. فقد يحتال أشرار بارعون على احتياطاتنا ويُطلقون آلات خارقة لا يُمكن السيطرة عليها وليس للبشرية قدرة على حماية نفسها منها. وإن نجونا من هذا، فقد نجد أنفسنا نضعُف تدريجيًّا مع نقل المزيد والمزيد من معرفتنا ومهاراتنا للآلات. قد تنصحُنا الآلات بعدم فعل هذا، لأنَّها تُدركُ القيمة الطويلة الأمد للاستقلالية البشرية، لكنَّنا قد نتجاهلُ نصائحها.

(١) الآلات النافعة

يقوم النموذجُ القياسيُّ الذي يعتمد عليه قدْر كبير من تقنيات القرن العشرين على آلاتٍ تسعى على النحو الأمثل لتحقيق هدفٍ ثابتٍ جرى تزويدُها به من الخارج. وكما رأينا، هذا النموذج بالأساس معيب. فهو ينجح فقط إن كان هناك ضمان بأنَّ الهدف كامل وصحيح، أو إن كان من السهولة بمكانٍ إيقاف الآلة. وهذان الشَّرطان لن يتحقَّقا مع اكتساب الذَّكاء الاصطناعي لمزيدٍ من الفاعلية والقوة.

إن كان من المُمكن أن يكون الهدف المزوَّد من الخارج خاطئًا، فمن غير المنطقي أن تتصرَّف الآلة وكأنه صحيح على الدَّوام. ومن هنا جاءت رُؤيتي للآلات النافعة: الآلات التي أفعالُها من المتوقَّع أن تُحقِّق أهدافنا «نحن». ولأنَّ تلك الأهداف موجودة بداخلنا وليس بداخل الآلات، فستحتاج الآلات إلى معرفة المزيد عما نرغبُ فيه بالفعل من ملاحظة الاختيارات التي نقوم بها وكيفية قيامنا بها. إن الآلات المُصمَّمة على هذا النحو ستكون خاضعةً للبشر؛ ستطلبُ الإذن منهم وستتصرَّف بحذَرٍ عندما تكون التوجيهات غير واضحة وستسمح بأن يُوقف تشغيلها.

في حين أن تلك النتائج الأولية خاصَّة بإعدادٍ مُبسَّط ومثالي، فأعتقد أنها ستستمرُّ عند التحوُّل إلى إعدادات أكثر واقعية. لقد طبَّق زملاء لي بالفعل بنجاحٍ نفس التوجُّه في التعامل مع مُشكلاتٍ عملية مثل تفاعُل السيارات الذاتية القيادة مع السائقين البشريِّين.[1] على سبيل المثال، من المعروف عن السيارات الذاتية القيادة أنها لا تُجيد التعامُل مع علامات التوقُّف الرباعي عندما لا يكون من الواضح من لدَيه الأولوية في المرور. لكن بصياغة ذلك في شكل لعبة تعاونية، تأتي السيارة بحلٍّ مُبتكر؛ إنها في واقع الأمر تتراجع إلى الخلف قليلًا لتُشير على نحوٍ واضح أنها لا تُخطِّط للسير أولًا. يفهم قائد السيارة تلك الإشارة ويسير إلى الأمام، وهو واثق بأنه لن يكون هناك تصادُم. من الواضح أننا — نحن الخبراء البشريين — كان بإمكاننا التفكير في هذا الحل وبرمجته في المركبة؛ لكن هذا لم يحدُث؛ فقد كان هذا نوعًا من التواصُل ابتكرتْه المركبة بنفسها بالكامل.

مع اكتسابنا لمزيدٍ من الخبرة من خلال إعداداتٍ أخرى، أتوقَّع أننا سنتفاجأ بنطاق وطلاقة سلوكيات الآلات عند تفاعُلها مع البشر. إننا مُعتادُون بشدة على غباء الآلات التي تُنفِّذ سلوكياتٍ مُبرمجة غير مرنة أو تسعى إلى تحقيق أهدافٍ مُحدَّدة، ولكنَّها غير صحيحة، والتي قد نُصدم من مدى المنطقية الذي أصبحت عليه. إن تقنية الآلات النافعة على نحوٍ مُثبت هي أساس توجُّهٍ جديدٍ للذَّكاء الاصطناعي ولبُّ علاقةٍ جديدةٍ بين البشر والآلات.

يبدو من المُمكن أيضًا تطبيق أفكارٍ مُماثلة فيما يتعلَّق بإعادة تصميم «الآلات» الأخرى التي من المُفترض أنها تخدُم البشر، بدءًا من النُّظُم البرمجية العادية. لقد تعلَّمنا كيفية إنشاء برمجيات بكتابة روتينات فرعية، كلٌّ منها لها «مواصفات» معروفة جيدًا تُحدِّد المخرجات التي ستنتج عن أحد المُدخلات؛ تمامًا كما هو الحال بالنِّسبة إلى زرِّ الجذر التربيعي في أيِّ آلةٍ حاسبة. تلك المُواصفات هي المُقابل المُباشر للهدف المُدمج في أيِّ نظام ذكاءٍ اصطناعي. ليس من المفترض من الروتين الفرعي أن يتوقَّف وينقُل التحكُّم

إلى الطبقات الأعلى في النظام البرمجي حتى يُنتج مخرجاتٍ تتوافق مع المواصفات. (هذا يجب أن يُذكِّرك بنظام الذكاء الاصطناعي الذي يستمرُّ في مسعاه الضيِّق الأُفُق إلى تحقيق الهدف المُعطى له.) سيتمثَّل النهج الأفضل في السماح بوجود عدم يقين في المواصفات. على سبيل المثال، يُعطى للروتين الفرعي، الذي يقوم بعملية حوسبة رياضية معقَّدة على نحوٍ مُخيف، حدٌّ خطأ يُحدِّد الدقة المطلوبة للإجابة، ويكون عليه إنتاجُ حلٍّ صحيح داخل نطاق حدِّ الخطأ هذا. في بعض الأحيان، قد يتطلَّب هذا أسابيع من العمل الحوسبي. بدلًا من ذلك، قد يكون من الأفضل أن تكون هناك دقَّة أقل فيما يتعلق بالخطأ المسموح به، بحيث يمكن أن يأتي الروتين الفرعي بعد ٢٠ ثانية ويقول: «لقد وجدتُ حلًّا بأن «هذا» جيد. فهل هذا يكفي أم تُريدني أن أستمرَّ؟» في بعض الحالات، قد يستمرُّ طرح السؤال طوال الطريق حتى المستوى الأعلى من النظام البرمجي بحيث يُمكن للمُستخدِم البشريِّ أن يُوفِّر مزيدًا من الإرشاد للنِّظام. وحينها ستُساعد الإجابات البشرية في تنقيح المواصفات في كل المستويات.

يُمكن تطبيق نفس نوع التفكير على كياناتٍ مثل الحكومات والشركات. تتضمن العيوب الواضحة في الحكومات إبداء اهتمامٍ شديد بالتفضيلات (المالية وكذلك السياسية) لمن هم في سُدَّة الحكم وإبداء اهتمام قليل جدًّا بتفضيلات المحكومين. من المفترض أن تنقل الانتخابات التفضيلات للحكومة، لكن يبدو أن لها نطاق عرضٍ صغيرًا على نحوٍ ملحوظ (مشابهًا بعض الشيء لبايت واحد من المعلومات كلَّ بضع سنوات) بالنسبة إلى مُهمة مُعقَّدة كهذه. في عددٍ كبير جدًّا من الدول، الحكومة ببساطة وسيلة لفرض مجموعة من الناس إرادتهم على الآخرين. أما الشركات، فتقوم بجُهودٍ أكبر لمعرفة تفضيلات العملاء، سواء من خلال أبحاث السوق أو التقييم المُباشر في شكل قرارات الشراء. على الجانب الآخر، إن صياغة التفضيلات البشرية من خلال الإعلان والمؤثِّرات الثقافية وحتى الإدمان الكيميائي تُعدُّ طريقةً مقبولة للقيام بالعمل.

(٢) حوكمة الذَّكاء الاصطناعي

للذَّكاء الاصطناعي القُدرة على إعادة تشكيل العالم، وتجبُ إدارة عملية إعادة التَّشكيل وتوجيهها بطريقةٍ ما. إن كان العددُ الهائل للمُبادرات الخاصَّة بتطوير حكومة فعَّالة للذَّكاء الاصطناعي مؤشِّرًا لنا، فنحن في وضعٍ مُمتاز. فعدد كبير من الجهات تُشكِّل معًا

هيئة أو مجلسًا أو لجنة دولية. لقد حدَّد المُنتدى الاقتصاديُّ العالميُّ حوالي ٣٠٠ مُحاولة مُنفصلة لتطوير مبادئ أخلاقية للذَّكاء الاصطناعي. ويُمكن النظر إلى صُندوق بريدي الإلكتروني باعتباره دعوةً واحدة طويلة لعقد مُنتدى قمةٍ عالمي عن مُستقبل الحوكمة الدولية للتأثيرات الثقافية والأخلاقية لتقنيات الذكاء الاصطناعي الناشئة.

هذا يختلف تمامًا عما حدث في مجال الطاقة النَّووية. فبعد الحرب العالمية الثانية، أمسكت الولايات المتَّحدة بكلِّ أوراق اللَّعب النووية في يدَيها. وفي عام ١٩٥٣، اقترح الرئيس الأمريكي دوايت أيزنهاور على الأمم المتَّحدة إنشاء هيئةٍ دولية لتنظيم استخدام التقنيات النووية. وفي عام ١٩٥٧، بدأت الوكالة الدولية للطاقة الذَّرية عملها، وهي تُعدُّ الجهة الدولية الوحيدة المُشرفة على التطوير الآمن والمُفيد للطاقة النووية.

في المقابل، تمتلك العديد من الأيدي أوراق اللعب الخاصَّة بالذكاء الاصطناعي. بالطبع، تُموِّل الولايات المتحدة والصين والاتحاد الأوروبي الكثير من الأبحاث المتعلقة بالذكاء الاصطناعي، لكن تقريبًا كلها تتمُّ خارج معامل وطنية آمنة. إن باحثي الذكاء الاصطناعي في الجامعات جزء من مُجتمع دوليٍّ واسع مُتعاون، يتلاحم أفرادُه معًا من خلال المصالح المشتركة والمؤتمرات واتفاقيات التعاون والجمعيات المهنية مثل جمعية النهوض بالذكاء الاصطناعي ومعهد مهندسي الكهرباء والإلكترونيات، والذي يتضمَّن عشرات الآلاف من الباحثين والممارسين في مجال الذكاء الاصطناعي. على الأرجح، غالبية الاستثمارات في البحث والتطوير في مجال الذكاء الاصطناعي تتمُّ الآن داخل الشركات، سواء الكبيرة منها أو الصغيرة؛ اللاعبُون الأبرز بحلول عام ٢٠١٩ هم جوجل (بما في ذلك ديب مايند) وفيسبوك وأمازون ومايكروسوفت وآي بي إم في الولايات المتحدة وتنسنت وبايدو، وإلى حدٍّ ما، علي بابا في الصين؛ وذلك ضمن كُبرى الشركات في العالم.[2] كل هذه الشركات فيما عدا تنسنت وعلي بابا أعضاء في مجموعة «الشراكة في الذكاء الاصطناعي»، وهي تحالُف صناعي يتضمَّن من بين مبادئه وعدًا بالتعاون فيما يتعلَّق بأمان الذَّكاء الاصطناعي. وأخيرًا، على الرغم من أن الغالبية العُظمى من البشر يمتلكُون القليل من الخبرة فيما يتعلَّق بالذكاء الاصطناعي، فهناك على الأقلِّ استعداد ظاهري فيما بين اللاعبين الآخرين لوضع مصالح البشر في الاعتبار.

هؤلاء، إذن، هم اللاعبون الذين يمتلكون غالبية أوراق اللعب في هذا المجال. إن مصالحهم لا تتوافق معًا على نحوٍ مثالي، لكنَّهم كلهم لديهم رغبة في السيطرة على نُظم الذكاء الاصطناعي عندما تُصبح أكثر قوة. (الأهداف الأخرى، مثل تجنُّب تفشي البطالة،

يشترك في تبنيها الحكومات والباحثون الجامعيون، ولكن ليس بالضَّرورة الشركات التي تتوقَّع التربُّح على المدى القصير من أكبر استخدامٍ مُمكن للذكاء الاصطناعي.) ولدعم هذا الاهتمام المُتبادل والقيام بتحرُّك مُتناسق، هناك مُنظَّمات لها «سلطة الدعوة إلى الاجتماعات»، وهذا يعني، على وجه التحديد، أنَّ المُنظَّمة إن نظَّمت اجتماعًا، فسيقبل الناسُ دعوة المشاركة فيه. فبالإضافة إلى الجمعيات المهنية، التي يُمكن أن تجمع باحثي الذَّكاء الاصطناعي معًا، ومجموعة «الشراكة في الذكاء الاصطناعي»، التي تجمع معًا الشركات والمعاهد غير الهادفة للرِّبح؛ فإنَّ الدعاة الأساسيِّين إلى الاجتماعات هم الأمم المتَّحدة (فيما يتعلَّق بالحكومات والباحثين) والمُنتدى الاقتصادي العالمي (فيما يتعلَّق بالحُكومات والشركات). وبالإضافة إلى ذلك، اقترحت مجموعة الدول الصناعية السبع الكُبرى إنشاء لجنة دولية معنية بالذكاء الاصطناعي، على أمل أن تكبُر وتُصبح يومًا شيئًا في حجم اللجنة الحكومية الدولية المعنية بتغيُّر المناخ التابعة للأمم المتحدة. إن التقارير الرنانة تتضاعف كما تتكاثرُ الأرانب.

في ظلِّ كل هذا النشاط، هل هناك احتمال لحُدوث تقدُّمٍ حقيقيٍّ فيما يتعلَّق بعملية الحوكمة؟ ما قد يدعُو إلى الدهشة أنَّ الإجابة هي نعم، على الأقل تدريجيًّا. إن العديد من الحكومات حول العالم تستعين بخدمات جهاتٍ استشارية لمساعدتها في عملية تطوير التشريعات؛ ربما المثال الأبرز هو مجموعة الخبراء الرفيعي المُستوى المعنية بالذَّكاء الاصطناعي التابعة للاتحاد الأوروبي. بدأت الاتفاقيات والقواعد والمعايير في الظهور فيما يتعلَّق بمسائل مثل خصوصية المُستخدمين وتبادُل البيانات وتجنُّب الانحياز العرقي. وتعمل الحكومات والشركات جاهدة من أجل الصياغة النهائية للقواعد الخاصة بالسيارات الذاتية القيادة؛ تلك القواعد التي لا محالة لها عناصرُ عابرة للحُدود. هناك إجماع على أنَّ القرارات الخاصَّة بالذكاء الاصطناعية يجب أن تكون قابلةً للتَّفسير حتى يُمكن الوثوق في نُظم الذَّكاء الاصطناعي، وهذا الإجماع قد تجلَّى بالفعل جزئيًّا في تشريع النظام العام لحماية البيانات الخاصِّ بالاتِّحاد الأوروبي. وفي كاليفورنيا، يحظُر قانون جديد أن تنتحل نظمُ الذَّكاء الاصطناعي شخصية البشر في ظروفٍ مُعيَّنة. هذان الأمران الأخيران — القابلية للتفسير والانتحال — بالتأكيد لهما بعض الأثر فيما يتعلَّق بمسألتي أمان الذكاء الاصطناعي والتحكُّم فيه.

في الوقت الحاضر، لا تُوجَد توصيات قابلة للتنفيذ يُمكن رفعها للحكومات أو غيرها من المؤسسات فيما يتعلَّق بمسألة الإبقاء على السيطرة على نُظم الذكاء الاصطناعي. إن

التشريع الذي يقول مثلًا: «يجب أن يكون نظام الذكاء الاصطناعي آمنًا وقابلًا للتحكم فيه» لن يكون له وزن؛ لأنَّ هذين المصطلحين ليس لهما حتى الآن معنًى دقيق، ولأنه لا تُوجَد منهجية هندسية معروفة على نطاقٍ واسع لضمان الأمان والقابلية للتفسير. لكن دعنا نكن متفائلين ونتخيل أنه بعد بضعة أعوام من العمل قد ثبتت صلاحية النهج المُتمثِّل في الذكاء الاصطناعي «النافع على نحو مُثبت» من خلال كلٍّ من التحليل الرياضي والتطبيق العملي في شكل تطبيقات مُفيدة. ربما، على سبيل المثال، يُصبح لدينا مُساعد رقمي شخصي يُمكننا الوثوق فيه، وجعلُه يستخدم بطاقات الائتمان الخاصة بنا ويُفرز مكالمتنا وبريدنا الإلكتروني، ويُدير أمورنا المالية؛ لأنه قد تكيَّف مع تفضيلاتنا البشرية وعرف متى يُمكنه المُضيُّ قُدُمًا بنفسه، ومتى من الأفضل أن يطلب مشورتنا. وربما تكون سياراتنا الذاتية القيادة قد تعلَّمت أُسس حُسن السلوك من أجل التفاعل بعضها مع بعض ومع السائقين البشريين، ومن المُفترَض أن تتفاعل الروبوتات المنزلية على نحوٍ سلس حتى مع أكثر الأطفال الصغار عنادًا. ومع وقوف الحظ في صفنا، لن يجري شوي أي قطط من أجل إعداد العشاء، ولن يجري تقديم لحم الحيتان لأعضاء حزب الخضر.

في تلك المرحلة، قد يكون من المُمكن تحديد قوالب التصميم البرمجي التي يجب أن تتوافق معها الأنواع المختلفة من التطبيقات حتى يجري بيعُها أو حتى تتصل بالإنترنت، تمامًا كما يجب على التطبيقات أن تمرَّ بعددٍ من الاختبارات البرمجية قبل أن يكون بالإمكان بيعُها على «أب ستور» الخاص بشركة أبل أو «جوجل بلاي». يستطيع مُصنِّعُو البرامج اقتراح قوالب إضافية، ما دام بإمكانهم تقديمُ براهين على أن القوالب تُلبِّي المتطلبات (التي ستكون حينها معرفة جيدًا) الخاصة بالأمان وقابلية التحكُّم. ستكون هناك آليات لإرسال تقارير بالأخطاء وتحديث النظم البرمجية التي تُنتج سلوكًا غير مرغوب فيه. وسيكون من المنطقي أيضًا إنشاء مدونات سلوك مهنية متعلقة بفكرة برامج الذكاء الاصطناعي النافعة على نحوٍ مُثبت ودمج الطرق والمُبرهنات المناظرة في المنهج الدراسي ذي الصلة من أجل إلهام الممارسين في مجال تعلُّم الآلة والذكاء الاصطناعي.

بالنسبة إلى مُراقب مُخضرم لوادي السيليكون، قد يبدو هذا ساذجًا بعض الشيء. فهناك تُوجَد معارضة شديدة لأي تشريعٍ من أيِّ نوع. وفي حين أننا مُعتادُون على فكرة أن شركات الأدوية يجب أن تُثبت الأمان والفاعلية (النافعة) لأي دواء من خلال التجارب الإكلينيكية قبل أن تُقدمه للعامة، فإن صناعة البرمجيات تعمل وفق مجموعةٍ مختلفة من القواعد؛ بعبارة أخرى، المجموعة الخالية. يُمكن «لمجموعة من المهندسين المتأنِّقين الذين

يرتشفُون بسرعة أحد مشروبات الطاقة»[3] في إحدى شركات البرمجيات إطلاق مُنتَجٍ أو تحديثٍ يؤثِّر تقريبًا على مليارات البشر دون وجود أي رقابة خارجية على الإطلاق.

لكن في النهاية سيكون على الصناعة التِّقنية أن تُدرك أن منتجاتها مُهمة، وما دامت منتجاتها كذلك، فمن المهم ألا تكون لها تأثيرات ضارَّة. هذا يعني أنه ستكون هناك قواعد تحكم طبيعة التفاعُل مع البشر وتحظر التصميمات التي، لنقُل، تتلاعب باستمرار بالتفضيلات أو تؤدي إلى سلوكٍ إدماني. أنا ليس لديَّ شكٌّ في أن التحوُّل من عالَم غير ذي قواعد إلى آخر ذي قواعد سيكون مُؤلمًا. دعنا نأمُل ألا يتطلَّب التغلُّب على مقاومة الصناعة حدوثَ كارثةٍ في حجم كارثة تشيرنوبل (أو ما هو أسوأ من هذا).

(٣) إساءة الاستخدام

إن تنظيم صناعة البرمجيات قد يكون أمرًا مُؤلمًا، لكنه لن يكون محتملًا بالنسبة إلى الأشرار الذين يُخطِّطون للهيمنة على العالم من أوكارهم السرية الموجودة تحت الأرض. لا شك أن العناصر الإجرامية والإرهابيين والأمم المارقة سيكون لديها دافع لتجنُّب وجود أيِّ قيود على تصميم الآلات الذكية حتى يُمكن استخدامها للتحكم في الأسلحة أو لابتكار أنشطةٍ إجرامية وتنفيذها. إن الخطر لا يكمُن في أن الخُطط الشرِّيرة سوف تنجح بقدْر ما أنه يتمثَّل في أنها ستفشل بسبب فقْد القُدرة على التحكُّم في النُّظم الذكية السيئة التصميم، وخاصَّة تلك المدمجة فيها أهداف شرِّيرة والمتاح لها استخدام أسلحة.

هذا ليس سببًا لتجنُّب القيام بعملية التنظيم؛ ففي النهاية، نحن لدَينا قوانين للقتل حتى وإن كان يجري التحايُل عليها في الغالب. لكن هذا يخلُق مُشكلةً مُهمَّةً جدًّا مُتعلِّقة بالمراقبة. إننا بالفعل نخسر معركتنا ضد البرامج الضارة والجرائم الإلكترونية. (يُقدِّر تقرير حديث عدد الضحايا في هذا الشأن بأكثر من ملياري شخص، والتكلفة السنوية بنحو ٦٠٠ مليار دولار.)[4] ستكون البرامج الضارة التي في شكل برامج عالية الذكاء أصعب كثيرًا في مواجهتها.

اقترح البعض، من بينهم نيك بوستروم، أن نستخدم نُظم الذكاء الاصطناعي الخارقة النافعة الخاصة بنا في اكتشاف أيِّ نُظم ذكاء اصطناعي ضارَّة أو سيئة السلوك على أيِّ نحوٍ آخر وتدميرها. بالتأكيد، يجبُ أن نستخدم الأدوات المتاحة أمامنا، مع تقليل تأثير ذلك على حريتنا الشخصية، لكنَّ صورة البشر الذين يحتشدُون في الأوكار، وهم يفتقدون

القُدرة على الدفاع عن أنفسهم ضدَّ القوات الهائلة التي تنتج عن مواجهة الآلات الخارقة، بالكاد مُطمئنة حتى لو كان بعضها في صفِّنا. سيكون من الأفضل كثيرًا إيجاد طرق لوأد الذكاء الاصطناعي الضار في المهد.

تتمثَّل أولى الخطوات الجيدة في إطلاق حملة ناجحة ومُتناسقة ودولية ضد الجرائم الإلكترونية، بما في ذلك توسيع نطاق اتفاقية بودابست المعنية بالجرائم الإلكترونية. سيُشكِّل هذا قالبًا تنظيميًّا للجهود المستقبلية المُمكنة لمنع ظهور برامج الذكاء الاصطناعي غير المُتحكَّم فيها. وفي نفس الوقت، سيُولِّد فهمًا ثقافيًّا واسعًا يرى أن إنشاء هذه البرامج، سواء عن قصد أو عن غير قصد، يُعدُّ على المدى الطويل بمنزلة عملٍ انتحاريٍّ يُقارَن بصنع كائنات وبائية.

(٤) الضعف واستقلالية البشر

استعرضت روايات إي إم فورستر الأكثر شُهرة، بما في ذلك «هاوردز إن» و«رحلة إلى الهند»، المجتمع البريطاني ونظامه الطبقي في الجزء الأول من القرن العشرين. في عام ١٩٠٩، كتب فورستر إحدى قصص الخيال العلمي البارزة، وهي «الآلة تتوقف». إن أهم ما يميز تلك القصة تبصُّرها، بما في ذلك تصويرها لـ (ما نُطلق عليه الآن) الإنترنت والمؤتمرات المرئية وأجهزة الآي باد والدورات الدراسية المفتوحة الواسعة النطاق عبر الإنترنت، وانتشار السِّمنة، وتجنُّب التواصُل المباشر. إن الآلة المذكورة في العنوان عبارة عن بِنية تحتية ذكية جامعة تفي بكل الاحتياجات البشرية. يُصبح البشر على نحوٍ مُتزايد مُعتمِدين عليها، لكنهم لا يعرفون كثيرًا عن كيفية عملها. إن المعرفة الهندسية تفسح المجال أمام ظهور تعاويذ طقسية تفشل في النهاية في وقف التدهور التدريجي لعمل الآلة. يرى كونو، الشخصية الرئيسية، ما يحدُث ولكنه لا يستطيع منعه:

> ألا يُمكنك أن ترَي ... أننا نحن من نموت وأن الآلة هي الشيء الوحيد الذي يحيا حقًّا هنا بالأسفل؟ لقد صنعنا الآلة كي تُنفِّذ إرادتنا، ولكنَّنا لا نملك أن ندفعها إلى تنفيذها الآن. لقد سلبتنا إحساسنا بالمكان وإحساسنا باللمس، وقد شوَّهت كل الصلات البشرية وشلَّت أجسادنا وإراداتنا. ... نحن موجودون فقط ككريات دمٍ تسري في شرايينها، وإذا كانت قادرة على العمل بدوننا، فسوف تتركنا نموت. أوه، أنا ليس لديَّ حل؛ أو لديَّ على الأقل حل واحد، والذي يتمثَّل

في إخبار الناس مرارًا وتكرارًا أنني رأيت تلال ويسيكس كما رآها ألفريد عندما أطاح بالدنماركيين.

لقد عاش أكثر من مائة مليار شخص على كوكب الأرض. وقد قضوا تقريبًا تريليون سنة يتعلَّمون ويُعلِّمون حتى يُمكن لحضارتنا أن تستمرَّ. وحتى الآن، الاحتمالية الوحيدة للاستمرار هي عن طريق إعادة الإنتاج في عقول الأجيال الجديدة. (إنَّ الورق يُعَدُّ وسيلة نقلٍ جيدة، ولكنه لا يفعل شيئًا حتى تصل المعرفة المسجَّلة عليه إلى عقل الشخص التالي.) هذا يتغير الآن؛ فعلى نحوٍ مُتزايد، من المُمكن أن ننقُل معرفتنا إلى الآلات التي يُمكنها بمفردها إدارة حضارتنا بالنيابة عنا.

بمجرَّد أن يختفي دافعنا العملي لتوريث حضارتنا للجيل التالي، سيكون من الصعب للغاية عكس العملية. وسيضيع فعليًّا تريليون سنة من التعلُّم المُتراكم. وسنُصبح ركَّابًا في باخرةٍ عملاقة تقودها الآلات، في رحلة مُستمرَّة للأبد؛ تمامًا كما هو مُتخيَّل في فيلم الرسوم المتحركة «وول-إي».

إن العواقبي الذكيَّ سيقولُ: «من الواضح أن تلك نتيجة غير مرغوب فيها للاستخدام المفرط للأتمتة! إن الآلات المُصمَّمة على نحوٍ مُلائم لن تفعل هذا أبدًا!» هذا صحيح، لكن فكر فيما يعنيه هذا. قد تدرك الآلات جيدًا أن الكفاءة والاستقلالية البشرية سِمتان مُهمَّتان للكيفية التي نفضل أن نعيش بها حياتنا. وقد تُصرُّ على أن يحتفظ البشر بتحكُّمهم في مصلحتهم الشخصية ومسئوليتهم عنها؛ بعبارةٍ أخرى، سترفُض الآلات فعل ذلك. لكن نحن البشر الكُسالى قصيري النظر قد نرفُض هذا. تُوجَد هنا مأساة مشاع؛ بالنسبة لكل فرد، قد يبدو من غير المُجدي الانهماك في سنواتٍ من التعلُّم المُضني لاكتساب معرفة ومهاراتٍ تمتلكُها الآلات بالفعل؛ لكن إن فكر الجميع بهذه الطريقة، فسيفقد الجنس البشريُّ على نحوٍ جماعي استقلاليته.

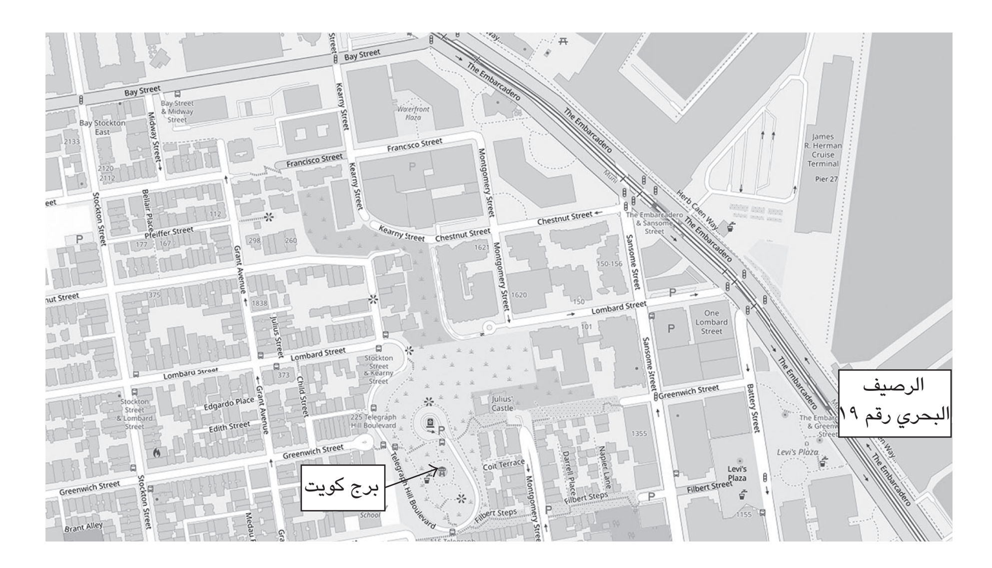
يبدو أن حلَّ هذه المشكلة ثقافي وليس تقنيًّا. سنحتاج إلى حركةٍ ثقافية لإعادة تشكيل مُثلنا وتفضيلاتنا باتجاه الاستقلالية والوساطة والقُدرة، وبعيدًا عن الترف والاعتمادية؛ إن شئتَ القول، نسخة ثقافية حديثة من الرُّوح العسكرية لإسبرطة القديمة. سيعني هذا هندسة التفضيلات البشرية على نطاقٍ عالمي إلى جانب إحداث تغييراتٍ جذرية في الطريقة التي يعمل بها مجتمعنا. ولتجنُّب جعل الوضع السيئ أسوأ، قد نحتاج إلى مساعدة الآلات الخارقة، من أجل تشكيل الحل وفي العملية الفعلية لتحقيق توازُن لكلِّ فرد.

إن هذه العملية مألوفة لأيِّ أبٍ لطفلٍ صغير. فبمُجرَّد أن يتجاوز الطفل المرحلة التي لا يستطيع فيها مساعدة نفسه، تحتاج الرعاية الأبوية إلى توازنٍ مُتطوِّر دائمًا بين فعل كل شيءٍ للطفل وتركه بالكامل لرغباته يفعل ما يريد. في مرحلة مُعينة، يدرك الطفل أن الأب قادر على نحوٍ تام على ربط رباط حذاء الطفل ولكنَّه يختار عدم فعل ذلك. هل هذا سيكون هو مُستقبل الجنس البشري؛ أي سيعامل كطفل، على الدوام، من جانب آلات تفوقه بشدة؟ أشكُّ في ذلك. أحد الأسباب هو أن الأطفال لا يُمكنُهم إيقاف آبائهم. (شكرًا للرب!) ولا يُمكننا أيضًا أن نُصبح حيوانات أليفة أو حيوانات تُودَع في حدائق الحيوان. لا يُوجَد حقًّا نظير في عالمنا الحالي للعلاقة التي ستكون بيننا وبين الآلات الذكية النافعة في المستقبل. سيكون علينا الانتظار لمعرفة كيف ستنتهي تلك المرحلة الختامية من اللعبة.

الملحق «أ»: البحث عن حلول

إن اختيار فعلٍ مُعيَّن بالاستباق ودراسة نتائج تسلسلات الأفعال الممكنة المختلفة يُعدُّ إحدى الإمكانيات الأساسية المتوفرة في النظم الذكية. إنه شيء يفعله هاتفك المحمول كلما سألته عن اتجاهاتٍ مُعينة. يعرض الشكل ١ مثالًا نموذجيًّا على ذلك؛ إذ يُوضِّح كيفية الانتقال من الموقع الحالي، الرصيف البحري رقم ١٩، إلى المكان المُستهدف وهو برج كويت. تحتاج الخوارزمية لمعرفة الأفعال المتاحة لها؛ عادةً، بالنِّسبة إلى تحديد المواقع باستخدام الخرائط، كل فعلٍ يجتاز قطاعًا من الطريق يربط بين تقاطُعَين مُتجاورَين. في المثال هنا، من الرصيف البحري رقم ١٩، هناك فعل واحد فقط؛ ألا وهو: الاتجاه يمينًا ثُم السير بطول طريق إمباركدرو حتى التقاطُع التالي. ثم هناك اختيار؛ وهو: الاستمرار أو الانعطاف الحاد نحو اليسار إلى شارع باتري. تستكشف الخوارزمية منهجيًّا كل الاحتمالات حتى تجد في النهاية طريقًا. إننا عادة ما نُضيف القليل من التوجيه المنطقي مثل تفضيل استكشاف الشوارع التي تتَّجه باتجاه المكان المستهدف وليس بعيدًا عنها. وبهذا التوجيه والقليل من الحيل الأخرى، يمكن للخوارزمية إيجاد حلولٍ مُثلى بسرعة جدًّا؛ عادة في ميلِّي ثوانٍ قليلة، حتى بالنِّسبة إلى رحلة عبر البلاد.

إن البحث عن مساراتٍ عبر الخرائط يُعدُّ مثالًا طبيعيًّا ومألوفًا، لكنه قد يكون مُضلِّلًا بعض الشيء لأن عدد الأماكن المميزة صغير للغاية. في الولايات المتحدة، على سبيل المثال، هناك فقط حوالي ١٠ ملايين تقاطُع. ربما يبدو هذا عددًا كبيرًا، لكنه صغير مقارنةً بعدد الأوضاع الأساسية في أحجية ١٥. إن أحجية ١٥ لعبة ذات إطار مساحته ٤ × ٤ يحتوي على ١٥ قطعة مُرقَّمة ومساحة واحدة فارغة. إن الهدف هو تحريك القطع لتحقيق هدفٍ مُعيَّن مثل ترتيب كل القطع على نحوٍ مُتسلسل رقميًّا. إن تلك الأحجية لها نحو ١٠ تريليونات

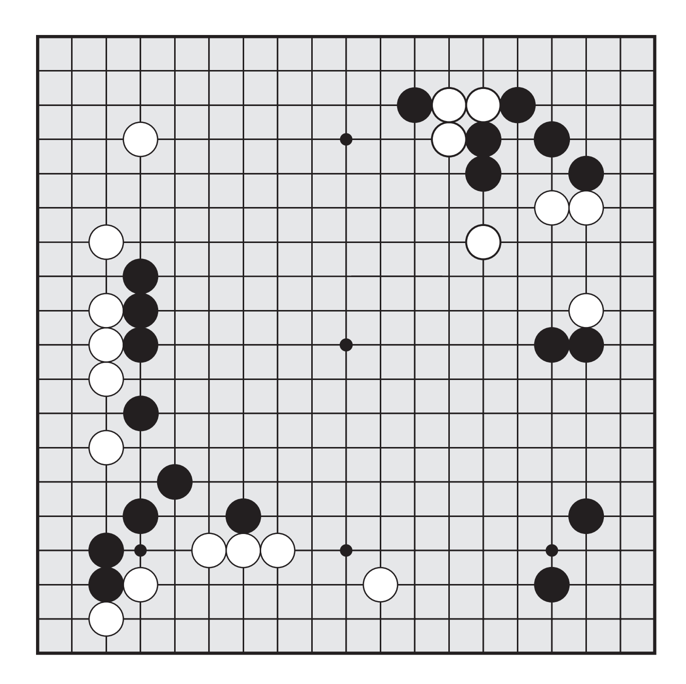


شكل ١: خريطة لجزء من سان فرانسيسكو تُوضِّح مكان الانطلاق والمُتمثل في الرصيف البحري رقم ١٩، والمكان المستهدف وهو برج كويت.

وضع (أي أكثر مليون مرة من عدد تقاطعات الولايات المتحدة!)، وللأحجية ٢٤ نحو ٨ تريليونات تريليون وضع. هذا مثال على ما يُطلق عليه علماء الرياضيات «التعقيد التوافقي»؛ أي الانفجار السريع لعدد التوافقيات مع زيادة عدد «الأجزاء المتحركة» لأي مشكلة. وبالعودة إلى مثال الولايات المتحدة، نجد أنه إن أرادت شركة نقل بالشاحنات تحسين تحرُّكات شاحناتها المائة عبر الولايات المتحدة، فإن عدد الأوضاع المُمكنة التي عليها وضعها في الاعتبار سيكون ١٠ ملايين أُس ١٠٠ (أي ١٠^٧٠٠).

(١) التخلِّي عن مُحاولة الوصول إلى قراراتٍ عقلانية

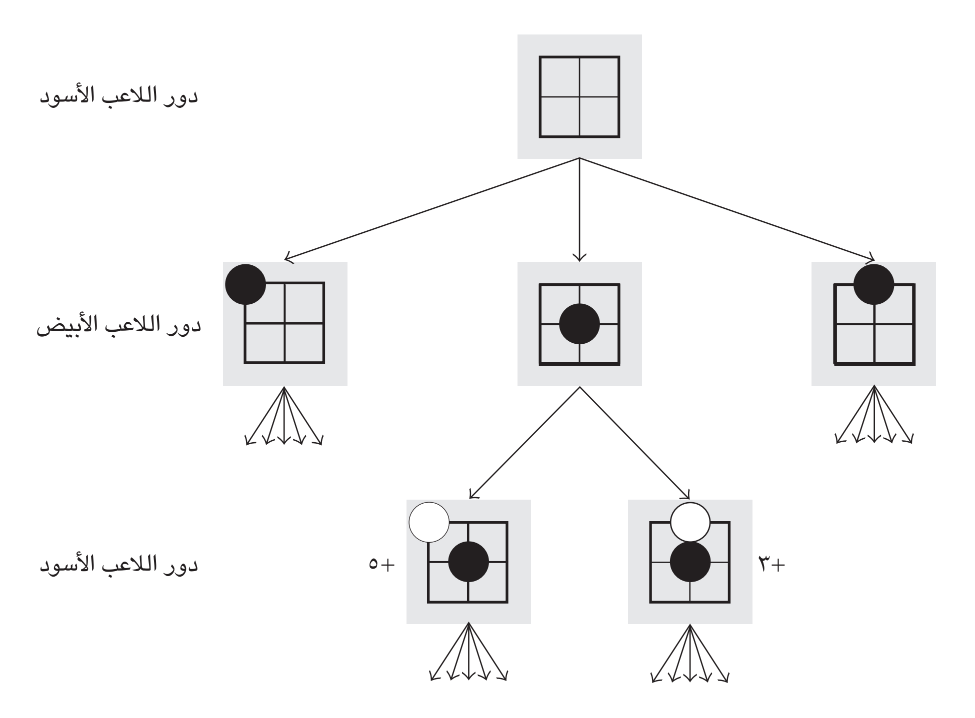
للعديد من الألعاب تلك الخاصية الخاصة بالتعقيد التوافُقي، بما في ذلك الشطرنج والدامة والطاولة ولعبة جو. ولأنَّ قواعد لعبة جو بسيطة ومُتميِّزة (انظر الشكل ٢)، سأستخدمُها كمثالٍ مُمتد. إن هدف اللعبة واضح بالقدر الكافي: تحقيق الفوز بالإحاطة بمساحة أكبر من خصمك. وتمامًا كما هو الحال فيما يتعلَّق بتحديد المواقع باستخدام



شكل ٢: لوح لعبة جو، أثناء المباراة الخامسة في نهائي كأس إل جي لعام ٢٠٠٢ بين ليو سيدول (اللاعب الأسود) وتشو مايونج-هون (اللاعب الأبيض). يتبادل اللاعبان وضع قطعةٍ واحدة في أيِّ مكان فارغ على اللوح. هنا، كان الدور على اللاعب الأسود للحركة وهناك ٣٤٣ حركة مُحتملة. يُحاول كل طرف إحاطة أكبر قدر مُمكن من المساحة. على سبيل المثال، اللاعب الأبيض لديه فرص جيدة لاكتساب مساحة في الحافة اليسرى وفي الجانب الأيسر من الحافة السُّفلية، في حين أن اللاعب الأسود قد يكتسب مساحةً في الركن الأيمن العُلوي والركن الأيمن السُّفلي. هناك مفهوم أساسي في هذه اللعبة وهو مفهوم «المجموعة»؛ أي مجموعة من القطع التي لها نفس اللون والمُرتبطة ببعضها من خلال تجاور رأسي أو أفقي. تبقى أيُّ مجموعة حيةً ما دامت هناك مساحة واحدة فارغة على الأقل بجوارها، أما إذا جرت إحاطتها بالكامل، مع عدم وجود أي مساحاتٍ فارغة، فستموت وتُزال من اللوح.

خريطة، فإنَّ الطريقة الواضحة لتحديد ماذا تفعل هو تخيُّل الأوضاع المُستقبلية التي ستنتج من تسلسلات الأفعال المُختلفة واختيار أفضلها. ستسأل: «إن فعلت هذا، ماذا قد يفعل خصمي؟ وماذا سأفعل حينها؟» تتَّضح تلك الفكرة في الشكل ٣ في لعبة جو ذات الإعداد ٣ × ٣. حتى في هذا الإعداد من اللعبة، يُمكنني عرض جزء صغير فقط من شجرة

الأوضاع المستقبلية المُمكنة، لكنَّني آمل أن تكون الفكرة واضحةً بالقدر الكافي. في واقع الأمر، هذه الطريقة في صنع القرارات تبدو بسيطة ومنطقية.



شكل ٣: جزء من شجرة اللعب الخاصة بلعبة جو ذات إعداد ٣ × ٣. بدءًا من الوضع الأوَّلي الخالي الذي يُطلق عليه «جذر» الشجرة، يُمكن للَّاعب الأسود اختيار واحدة من ثلاث حركات أساسية مُمكنة. (الحركات الأخرى مُتوافقة مع هذه الحركات.) وبعدها سيكون على اللاعب الأبيض الدور في التحرُّك. إن اختار اللاعب الأسود اللعب في المنتصف، فسيكون لدى اللاعب الأبيض حركتان أساسيتان — الركن أو الجانب — ثم سيكون على اللاعب الأسود اللعب ثانيةً. وبتخيُّل الأوضاع المُحتملة هذه، يمكن للَّاعب الأسود اختيار الحركة التي سيلعبها في الوضع الأوَّلي. إن لم يكن اللاعب الأسود قادرًا على تتبُّع كل خط لعب مُمكن حتى نهاية اللعبة، فيُمكن استخدام دالة تقييم لتقدير مدى جودة الأوضاع في أوراق الشجرة. هنا، تعين دالة التقييم +٥ و+٣ لاثنتين من الأوراق.

تتمثَّل المُشكلة في أن لعبة جو بها أكثر من ١٠^١٧٠ وضع مُحتمل في اللوح الكامل ذي الإعداد ١٩ × ١٩. وفي حين أن إيجاد أقصر مسارٍ مضمون على خريطةٍ سهل نسبيًّا، فإن إيجاد طريقة مضمونة للفوز في لعبة جو مُتعذِّر تمامًا. وحتى لو استكشفت الخوارزميةُ اللعبة للمليار عام القادمة، فيُمكنها استكشاف قدْرٍ بسيط فقط من شجرة الاحتمالات بأكملها. يؤدي بنا هذا إلى سؤالَين. الأول هو: أي جزء من الشجرة يجب أن يستكشفه البرنامج؟ والثاني هو: أي حركةٍ يجب على البرنامج أن يقوم بها، في ضوء جزء الشجرة الذي استكشفه؟

للإجابة عن السؤال الثاني، الفكرة الأساسية التي تستخدمها تقريبًا كل البرامج الاستباقية هي تعيين «قيمة تقديرية» لـ «أوراق» الشجرة — تلك الأوضاع الأبعد في المُستقبل — ثم العمل من أجل تحديد مدى فاعلية الاختيارات عند الجذر.[1] على سبيل المثال، بالنظر إلى الوضعَين في الجزء السُّفلي من الشكل ٣، قد يُخمِّن المرء قيمةً قدرها +٥ (من وجهة نظر اللاعب الأسود) للوضع الذي على اليسار و+٣ للوضع الذي على اليمين؛ لأن قطعة لعب اللاعب الأبيض في الركن مُعرَّضة للخطر أكثر من تلك التي على الجانب. إن كانت هاتان القيمتان صحيحتَين، فيُمكن أن يتوقع اللاعب الأسود أن اللاعب الأبيض سيلعب على الجانب، مما يؤدي إلى الوضع الأيمن؛ ومن ثم، يبدو من المعقول تعيين قيمة +٣ للحركة الأولية للاعب الأسود في المُنتصف. ومع بعض التغييرات البسيطة، تعدُّ هذه هي الخطة التي استخدمها برنامج لعب الدامة الذي صمَّمه آرثر صمويل لهزيمة مُصمِّمه في عام ١٩٥٥،[2] و«ديب بلو» لهزيمة بطل العالم حينها في لعبة الشطرنج، جاري كسبروف، في عام ١٩٩٧، و«ألفا جو» لهزيمة بطل العالم السابق في لعبة جو لي سيدول في عام ٢٠١٦. بالنسبة إلى جهاز «ديب بلو»، كتب البشر جزء البرنامج الذي قيَّم الأوضاع التي عند أوراق الشجرة، على نحوٍ كبير بناءً على معرفتهم بلعبة الشطرنج. بالنسبة إلى برنامج صمويل وبرنامج «ألفا جو»، فقد تعلَّما ذلك من آلاف أو ملايين المباريات التجريبية.

السؤال الأول — أيُّ جزء من الشجرة يجب أن يستكشفه البرنامج؟ — مثال على أحد أهمِّ الأسئلة في مجال الذكاء الاصطناعي؛ ألا وهو: «ما عمليات الحوسبة التي يجب على أيِّ كيان ذكي القيام بها؟» بالنسبة إلى برامج لعب الألعاب، إنه يُعَدُّ سؤالًا مُهمًّا جدًّا؛ لأن لتلك البرامج نطاقًا زمنيًّا صغيرًا وثابتًا، واستهلاكه في القيام بعمليات حوسبة لا قيمة لها طريقة أكيدة للخسارة. وبالنسبة إلى البشر والكيانات الأخرى التي تعمل في العالم

الواقعي، إنه مُهمٌّ أكثر لأن العالم الواقعي أعقد بكثيرٍ جدًّا؛ فما لم يُحدَّد قدر الحوسبة المطلوب بعناية، لن يستطيع أيُّ قدرٍ من الحوسبة القيام بأيِّ دورٍ في حلِّ مشكلة تحديد ما يجبُ فعلُه. إذا كنتَ تقود سيارتك وحيوانُ موظ يسير في مُنتصف الطريق، فلا فائدة من التفكير فيما إذا كان يجب استبدال اليوروهات بالجنيهات أو ما إذا كان على اللاعب الأسود أن يجعل حركته الأولى في مُنتصف لوح لعبة جو.

إن قدرة البشر على إدارة نشاطهم الحوسبي بحيث تُتَّخذ قرارات معقولة بسرعة معقولة على الأقل ملحوظة مثل قُدرتهم على الإدراك والتفكير على نحوٍ صحيح. ويبدو أنها شيء نكتسبه على نحوٍ طبيعي ودون جهد؛ فعندما علَّمني أبي لعب الشطرنج، علَّمني القواعد، لكنه لم يُعلمني الخوارزمية الجيِّدة الخاصة باختيار أجزاء شجرة اللعبة التي يجب استكشافها، وتلك التي يجب تجاهلها.

كيف يحدث هذا؟ وعلى أيِّ أساسٍ يُمكننا توجيه أفكارنا؟ تتمثَّل الإجابة في أن أي عملية حوسبة لها قيمة مُتعلِّقة بمدى تحسينها لنوعية قرارك. إن عملية اختيار عمليات الحوسبة تُسمَّى «ما وراء التفكير»، والتي تعني التفكير في التفكير. وكما أن الأفعال يمكن أن تُختار بعقلانية، على أساس القيمة المتوقعة، فيمكن أن يحدث نفس الشيء مع عمليات الحوسبة. ويُطلق على هذا «ما وراء التفكير العقلاني».[3] إن الفكرة الأساسية هنا بسيطة جدًّا:

> هل عمليات الحوسبة ستُقدِّم أعلى تحسين مُتوقَّع لنوعية القرار وستتوقَّف عندما تتجاوز التَّكلفة (فيما يتعلق بالوقت) التحسن المُتوقَّع؟

هذا هو كل شيء. لا حاجة إلى خوارزمية مُعقَّدة! هذا المبدأ البسيط يُنتج سلوكًا حوسبيًّا فعالًا في نطاقٍ واسع من المشكلات، بما في ذلك لُعبتا الشطرنج وجو. ويبدو من المُحتمل أن أدمغتنا تُنفِّذ شيئًا مماثلًا، والذي يفسر السبب وراء عدم حاجتنا إلى تعلُّم خوارزميات جديدة ومتعلقة باللعبة للتفكير مع كل لعبة جديدة نتعلم لعبها.

إن استكشاف شجرة من الاحتمالات التي تمتدُّ إلى الأمام في المستقبل من الوضع الحالي لا يُعَد الطريقة الوحيدة للوصول إلى قرارات في واقع الأمر. عادةً، يكون أكثر منطقية العمل على نحوٍ عكسي من الهدف. على سبيل المثال، إن وجود حيوان الموظ في الطريق يقترح هدف: «تجنب الاصطدام بحيوان الموظ»، والذي بدوره يقترح ثلاثة أفعال مُمكنة؛ الانحراف يسارًا، أو الانحراف يمينًا، أو الضغط بقوة على المكابح. إنه لا يقترح

فعل مبادلة اليوروهات بالجنيهات أو وضع قطعة لعب سوداء في المنتصف. ومن ثمَّ، الأهداف لها تأثير تركيزي رائع على تفكير المرء. لا تستفيد أي برامج حالية خاصَّة بلُعب الألعاب من هذه الفكرة؛ في واقع الأمر، إنها عادة ما تتدبَّر كل الأفعال المُمكنة والمسموح بها. وهذا يُعَدُّ أحد الأسباب (العديدة) لعدم قلقي من سيطرة إصدار برنامج «ألفا جو» الذي يُسمَّى «ألفا زيرو» على العالم.

(٢) الاستباق على نحوٍ أكبر

دعنا نفترض أنك قررتَ القيام بحركةٍ معيَّنة على لوح لعبة جو. هذا أمر رائع! والآن، عليك أن تقوم بهذا بالفعل. في العالم الواقعي، يتضمَّن هذا مدَّ يدك داخل وعاء قطع اللعب التي لم تُستخدم بعدُ لالتقاط واحدةٍ منها، ثم تحريك يدك فوق المكان المراد ثم وضع القطعة ببراعة على الموضع إما بهدوء أو بقوة وفقًا لتقليد اللعبة.

إن كلًّا من هذه المراحل، بدوره، يتكوَّن من مجموعة مُعقَّدة من أوامر التحكم الحركي والمعرفي التي تتضمَّن العضلات والأعصاب الخاصة باليد والذراع والكتف والعينين. وبينما تمدُّ يدك لتصل إلى قطعة لعب، فأنت تتأكد من أنَّ بقية جسمك لن ينقلب بسبب التغيُّر في مركز الجاذبية الخاص بك. إن حقيقة أنك قد لا تكون مُدركًا على نحوٍ واعٍ لاختيارك لتلك الأفعال لا يعني أن دماغك لم تخترْها. على سبيل المثال: ربما تكون هناك العديد من قطع اللعب في الوعاء، لكن «يدك» — في واقع الأمر، دماغك الذي يُعالج المعلومات الحسية — لا يزال عليه اختيار إحداها كي يجري التقاطها.

تقريبًا كل شيءٍ نفعله يُشبه هذا. ففي أثناء قيادة السيارة، قد نختار «الانتقال إلى الحارة اليُسرى من الطريق»، لكن هذا الفعل يتضمَّن النظر في المرآة وفوق كتفك وربما تعديل السرعة وتحريك عجلة القيادة مع مراقبة التقدُّم حتى يتم الأمر بنجاح. في الحوارات، يتضمَّن أيُّ ردٍّ روتيني مثل: «حسنًا، دعني أُراجع دفتر مواعيدي وأعود إليك» نُطق العديد من المقاطع الصوتية التي يتطلَّب كل منها مئات أوامر التحكُّم الحركي المتناسقة على نحوٍ دقيق لعضلات اللسان والشفتَين والفك والحلق والجهاز التنفُّسي. بالنسبة إلى لغتك الأم، هذه العملية آلية؛ إنها تُشبه كثيرًا فكرة تشغيل روتينٍ فرعيٍّ في برنامج كمبيوتر (ارجع إلى الفصل الثاني). إن حقيقة أن تسلسُلات الأفعال المعقَّدة يُمكن أن تُصبح روتينية وآلية؛ ومن ثمَّ تعمل بمنزلة أفعالٍ فردية في عمليات أكثر تعقيدًا، تُعدُّ

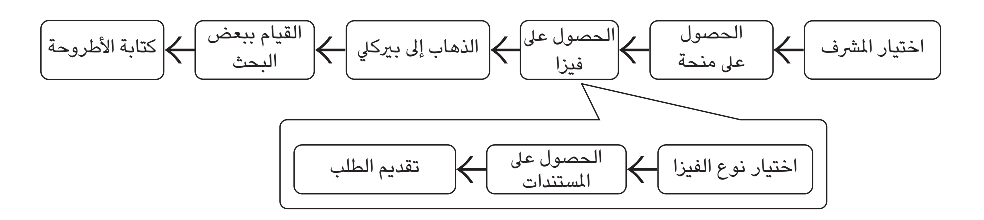
جوهريةً تمامًا للإدراك البشري. إن نُطق كلماتٍ في لغة غير شائعة — ربما السؤال عن كيفية الوصول لمدينة شتبجيشن في بولندا — يُعدُّ تذكرةً مُفيدة بأنه كان هناك وقت في حياتك كانت قراءة الكلمات ونُطقها مُهمَّتَين صعبتَين تتطلَّبان جهدًا ذهنيًّا ومُمارسةً كبيرة.

ومن ثمَّ، فالمشكلة الحقيقية التي يُواجهها دماغك لا تتمثَّل في اختيار القيام بإحدى الحركات على لوح لعبة جو، وإنما إرسال أوامر تحكُّم حركي لعضلاتك. وإذا حوَّلنا انتباهنا من مستوى حركات لعبة جو إلى مستوى أوامر التحكُّم الحركي، فستبدو المشكلة مُختلفة للغاية. بوجهٍ عام، يُمكن أن يُرسِل دماغك أوامر كل مائة ميلِّي ثانية تقريبًا. ونحن لدَينا نحو ٦٠٠ عضلة، ومن ثم، هناك حد أقصى نظري يُقدَّر بنحو ٦٠٠٠ أمر في الثانية، وعشرين مليونًا في الساعة، و٢٠٠ مليارٍ في السنة، و٢٠ تريليونًا على مدار العمر. عليك استخدامها بحكمة!

والآن، افترض أننا حاولنا تطبيق خوارزمية شبيهة بتلك الخاصة بإصدار برنامج «ألفا جو» المُسمى «ألفا زيرو» لحلِّ مشكلة اتخاذ القرار في هذا المستوى. في لعبة جو، يقوم هذا الإصدار بالاستباق ربما لخمسين خطوة. لكن خمسين خطوةً من أوامر التحكُّم الحركي تأخذك إلى بضع ثوانٍ فقط في المستقبل! وهذا ليس كافيًا للعشرين مليون أمر تحكُّمٍ حركي في مباراة للعبة جو التي تستمرُّ لمدة ساعة، وبالتأكيد غير كافٍ للخُطوات التريليون (١٠٠٠٠٠٠٠٠٠٠٠٠) المتضمَّنة في إعداد رسالة دكتوراه. ومن ثمَّ، حتى على الرغم من أن هذا الإصدار يقوم بالاستباق على نحوٍ أكبر في لعبة جو مما هو متاح لأي إنسان، فلا يبدو أن تلك القدرة مفيدة في العالم الواقعي. إنها النوع الخاطئ من الاستباق.

أنا لا أقصد بالطبع أن إعداد رسالة دكتوراه يتطلَّب فعليًّا التخطيط الجيد لتريليون خطوة عضلية مقدمًا. إن الخطط المجردة إلى حدٍّ كبير فقط هي التي تتمُّ في البداية؛ ربما اختيار جامعة كاليفورنيا ببيركلي أو مكانٍ آخر واختيار مُشرف على الرسالة أو موضوع البحث والتقدُّم من أجل الحصول على منحة والحصول على فيزا خاصَّة بالطلبة والسفر إلى المدينة المرادة والقيام ببعض البحث وغير ذلك. وللقيام باختياراتك، إنك تقوم بالقدر الكافي فقط من التفكير بشأن الأشياء الصحيحة فقط حتى يُصبح القرار واضحًا. إن كانت إمكانية إحدى الخطوات المجردة مثل الحصول على الفيزا غير واضحة، فستقوم بالمزيد من التفكير وربما بالمزيد من جمع المعلومات، مما يعني جعل الخطة ماديةً أكثر في بعض الجوانب؛ ربما اختيار نوع الفيزا الملائمة وتجهيز المستندات الضرورية وتقديم

الطلب. يعرض الشكل ٤ الخطة المجردة وتنقيح خطوة الحصول على الفيزا في خطةٍ فرعية ثلاثية الخطوات. وعندما يحين وقت البدء في تنفيذ الخطة، يجب تنقيح خطواتها المبدئية طوال المُستوى الأوَّلي حتى يُمكن لجسمك تنفيذها.



شكل ٤: خطة مجردة لطالب أجنبي اختار الحصول على رسالة الدكتوراه في جامعة كاليفورنيا ببيركلي. جرى توسيعُ خُطوة الحصول على الفيزا، التي إمكانيتُها غير مؤكَّدة، في خطة مجرَّدة خاصَّة بها.

إنَّ برنامج «ألفا جو» ببساطة لا يُمكنه القيام بهذا النوع من التفكير؛ إن الأفعال الوحيدة التي يضعها في اعتباره هي الأفعال الأوَّلية التي تحدُث في تسلسل من الحالة المبدئية. إنه ليس لديه مفهوم «الخطة المجردة». إن محاولة تطبيق طريقة تفكير برنامج «ألفا جو» في العالم الواقعي تُشبه محاولة كتابة رواية بالتساؤل عما إذا كان الحرف الأول يجب أن يكون «أ» أم «ب» أم «ج»، وهكذا.

في عام ١٩٦٢، أكد هربرت سيمن على أهمية التنظيم التسلسُلي في بحثٍ شهيرٍ بعُنوان «بنية التعقيد».[4] وطوَّر باحثُو الذكاء الاصطناعي منذ أوائل سبعينيات القرن الماضي مجموعةً متنوعة من الطرُق التي تُنشئ وتُنقِّح خططًا منظمة تسلسليًّا.[5] إن بعض النظم الناتجة قادرة على إنشاء خطط لها عشرات الملايين من الخطوات؛ على سبيل المثال، لتنظيم الأنشطة التصنيعية في مصنع كبير.

نحن الآن لدَينا فهم نظري جيد جدًّا لمعنى الأفعال المجرَّدة؛ أي لكيفية تعريف تأثيراتها على العالم.[6] تأمل، على سبيل المثال، الفعل المجرد «الذهاب إلى بيركلي» في الشكل ٤. إنه يمكن تنفيذه بطرُق عديدة مختلفة، التي لكلٍّ منها تأثيرات مختلفة على العالم: يمكن أن تذهب إلى هناك بحرًا أو تُسافر خلسة على متن سفينةٍ أو تطير إلى كندا وتعبر

الحدود من هناك أو تستأجر طائرة خاصة أو غير ذلك. لكنك لست بحاجة إلى التفكير في أيٍّ من تلك الاختيارات في الوقت الحاضر. فما دمتَ مُتأكِّدًا أن هناك طريقةً للقيام بالأمر لا تستهلك الكثير من الوقت والمال أو لها مخاطر كبيرة بحيث تُهدد بقية الخطة، فيمكنك فقط وضع تلك الخطوة المجردة في الخطة والاطمئنان بأن الخطة ستنجح. بهذه الطريقة، يمكنك إنشاء خططٍ عالية المستوى تتحوَّل في النهاية إلى مليارات أو تريليونات الخطوات الأولية دون القلق بشأن ماهية تلك الخطوات حتى يحين وقت تنفيذها الفعلي.

في واقع الأمر، ليس أيٌّ من هذا مُمكنًا بدون التسلسل. فبدون الأفعال العالية المستوى مثل الحصول على فيزا وكتابة أطروحة، لا يُمكننا إنشاء خطة مجرَّدة للحصول على رسالة الدكتوراه؛ وبدون الأفعال الأعلى مُستوًى مثل الحصول على الدكتوراه وإنشاء شركة، لا يمكننا التخطيط للحصول على الدكتوراه ثم إنشاء شركة. في العالم الواقعي، سنفشل إن لم تكن لدينا مجموعة هائلة من الأفعال على عشرات المستويات من التجريد. (في لعبة جو، لا يوجد تسلسل واضح للأفعال، لذا مُعظمُنا يخسر.) لكن في الوقت الحاضر كل الطرق الموجودة للتخطيط التسلسلي تعتمد على تسلسلٍ أنتجه الإنسان للأفعال المجرَّدة والمادية؛ فنحن لم نفهم بعد كيف يُمكن تعلُّم تلك التسلسلات من خلال التجربة.

الملحق «ب»: المعرفة والمنطق

المنطق هو دراسة التفكير في معرفة معينة. إنه عام على نحوٍ تامٍّ فيما يتعلَّق بالموضوع؛ أي المعرفة يُمكن أن تكون مُتعلِّقة بأي شيء. ومن ثم فالمنطق يُعدُّ جزءًا لا غِنى عنه من فهمنا للذكاء العام.

إن المتطلب الأساسي للمنطق هو لغة «صورية» ذات معانٍ دقيقة للجُمَل التي في اللغة، بحيث تُوجَد عملية واضحة لتحديد ما إذا كانت إحدى الجُمَل صحيحةً أم خاطئة في موقفٍ مُعيَّن. هذا هو كل شيء. وبمُجرَّد أن يكون لدينا هذا، يُمكننا كتابة خوارزميات تفكير «جيد» تُنتج جُمَلًا جديدة من جُملٍ معروفة بالفعل. تلك الجمل الجديدة تنبع بالتأكيد من الجمل التي يعرفها النظام بالفعل، بمعنى أن الجمل الجديدة تكون صحيحة بالضرورة في أيِّ موقفٍ تكون فيه الجُمَل الأصلية صحيحة. يسمح هذا لأيِّ آلة بالإجابة عن أسئلة أو إثبات مُبرهنات رياضية أو إنشاء خُطط مضمون نجاحها.

إنَّ جبر المرحلة الثانوية يقدم مثالًا جيدًا (على الرغم من أنه قد يجعلنا نتذكَّر ذكريات مؤلمة). تتضمن اللغة الصورية جملًا مثل ٤س + ١ = ٢ص − ٥. هذه الجملة صحيحة في الوضع الذي يكون فيه س = ٥ وص = ١٣، وخاطئة في الوضع الذي يكون فيه س = ٥ وص = ٦. من تلك الجملة، يُمكن استنباط جملة أخرى مثل ص = ٢س + ٣، وعندما تكون الجملة الأولى صحيحة، يجب أن تكون الثانية كذلك أيضًا.

إنَّ الفكرة الأساسية للمنطق، التي جرى تطويرُها على نحوٍ مُنفصل في اليونان والصين والهند القديمة، تتمثَّل في أن نفس المفاهيم الخاصَّة بالمعنى الدقيق والتفكير السليم يُمكن تطبيقها على جمل تتعلَّق بأيِّ مجال، وليس على الأعداد فقط. المثال القياسي يبدأ بالآتي: «سقراط رجل» و«كل الرجال فانون» وينتهي إلى ما يلي: «سقراط فانٍ».[1]

هذا الاستنباط صوري تمامًا بمعنى أنه لا يعتمد على أيٍّ معلوماتٍ أخرى متعلِّقة بماهية سقراط أو معنى كلمتي «رجل» و«فانٍ». إن حقيقة أن التفكير المنطقي صوري تمامًا تعني أنه يُمكن كتابة خوارزميات لتنفيذه.

(١) منطق القضايا

لأغراضنا المُتعلِّقة بفهم إمكانات وآفاق الذكاء الاصطناعي، نرى أن هناك نوعَين مُهمَّين من المنطق: منطق القضايا، والمنطق الإسنادي. والفرق بين الاثنين جوهري لفهم الوضع الحالي للذكاء الاصطناعي والكيفية التي من المُنتظر أن يتطور بها.

دعنا نبدأ بمنطق القضايا، والذي هو أبسط من النوع الآخر. الجُمل تتكوَّن من نوعَين فقط من الأشياء: الرموز المُمثِّلة للقضايا التي قد تكون صحيحة أو خاطئة، و«الروابط» المنطقية مثل and (و) وor (أو) وnot (ليس) وif ... then (إذا كان ... فإن ...). (سنعرض لمثال بعد وقتٍ قصير.) تُسمَّى تلك الروابط المنطقية أحيانًا بالروابط «البولينية»، نسبة إلى جورج بول، وهو عالم منطق ينتمي إلى القرن التاسع عشر أعاد الحياة إلى المجال بأفكاره الرياضية الجديدة. إنها مُماثلة تمامًا «للبوابات المنطقية» المُستخدمة في رقاقات الكمبيوتر.

عُرفت الخوارزمياتُ العمليةُ الخاصة بالتفكير باستخدام منطق القضايا منذ أوائل ستينيات القرن الماضي.[2,3] وعلى الرغم من أن مهمة التفكير العام قد تتطلَّب وقتًا أُسيًّا في أسوأ الحالات،[4] فإن خوارزميات التفكير الحديث باستخدام منطق القضايا تُعالج مشكلات لها ملايين رموز القضايا وعشرات ملايين الجمل. إنها تُعدُّ أداةً أساسية لإنشاء خططٍ منطقية مضمونة والتحقق من تصميمات الرقاقات قبل تصنيعها والتأكد من صحة التطبيقات البرمجية وبروتوكولات الأمن قبل استخدامها. الشيء المذهل هو أن خوارزمية واحدة — خوارزمية تفكير يقوم على منطق القضايا — تحلُّ «كل» هذه المهام بمجرد صياغة تلك المهام على شكل مهامِّ تفكير. من الواضح أن تلك خطوة باتجاه غاية العمومية في النظم الذكية.

لسوء الحظ، هذه ليست خطوة كبيرة جدًّا لأنَّ لغة منطق القضايا ليست غالية جدًّا. دعنا نرى ما يعنيه هذا في الممارسة عندما نُحاول التعبير عن القاعدة الأساسية للحركات المسموح بها في لعبة جو: «يستطيع اللاعب الذي عليه الدور في اللعب وضْع قطعة اللعب على أي تقاطع خالٍ».[5] الخطوة الأولى تتمثَّل في تحديد رموز القضية التي

ستُستخدم في الحديث عن حركات اللعب والأوضاع على اللوح. إن القضية الأساسية المُهمَّة هي ما إذا كانت قطعةُ اللعب التي من لون معين موجودة في موضع مُعيَّن في وقتٍ مُعيَّن. ومن ثم، نحتاج إلى رموزٍ مثل القطعة_البيضاء_على_٥_٥_في_الحركة_٣٨، والقطعة_السوداء_على_٥_٥_في_الحركة_٣٨. (كما هو الحال مع كلمات «رجل» و«فانٍ» و«سقراط»، تذكَّر أن خوارزمية التفكير لا تحتاج إلى معرفة معنى الرموز.) ومن ثمَّ فإن الشرط المنطقي لقطعة اللعب البيضاء حتى تكون قادرةً على الانتقال إلى تقاطع ٥،٥ في الحركة ٣٨ سيكون:

(ليست القطعة_البيضاء_على_٥_٥_في_ الحركة_٣٨)

و(ليست القطعة_السوداء_على_٥_٥_في_الحركة_٣٨)

بعبارة أخرى، لا تُوجَد قطعة لعب بيضاء ولا سوداء. يبدو هذا بسيطًا بالقدر الكافي. لكن لسوء الحظ، في منطق القضايا، يجب كتابة هذا على نحوٍ مُنفصل لكل موضع ولكل حركة في اللعبة. ولأنَّ هناك ٣٦١ موضعًا ونحو ٣٠٠ حركة في كلِّ مباراة، فهذا يعني أكثر من مائة ألف نسخة من القاعدة! وبالنسبة إلى القواعد الخاصة بالاستحواذ والتَّكرار، التي تتضمَّن قطع لعبٍ ومواضع مُتعدِّدة، الوضع أسوأ وسنملأ بسُرعة ملايين الصفحات.

إنَّ العالم الواقعي، على نحوٍ واضح، أكبر بكثيرٍ من لوح لعبة جو؛ هناك عدد أكبر بكثير جدًّا من الـ ٣٦١ موضعًا والخُطوات الزمنية الثلاثمائة، وهناك أنواع عديدة من الأشياء بجانب قطع اللعب، لذا، فإنَّ احتمال استخدام لغةٍ تقُوم على منطق القضايا للمعرفة الخاصة بالعالم الواقعي مُستبعد تمامًا.

إنَّ «الحجم» السخيف لكتاب القواعد ليس فقط هو المُشكلة؛ وإنَّما أيضًا قدرُ «التجربة» السخيف الذي سيحتاجه أيُّ نظام تعلُّمٍ لتعلُّم القواعد من الأمثلة. وفي حين أن الإنسان يحتاج فقط مثالًا أو مثالَين لمعرفة الأفكار الأساسية المُتعلِّقة بوضع قطعة اللعب والاستحواذ على قطع اللعب وما إلى ذلك، فيجب أن يُقدم لأي نظام ذكي يعتمد على منطق القضايا أمثلة على التحريك والاستحواذ على نحوٍ مُنفصل لكلِّ موضع وخُطوة زمنية. إن النظام ليس بإمكانه التعميم من مجرَّد بضعة أمثلة، كما يفعل الإنسان، لأنه ليست لديه طريقة للتعبير عن القاعدة العامَّة. وهذا القصور ينطبق ليس فقط على النُّظم القائمة على منطق القضايا، وإنَّما أيضًا على أيِّ نظامٍ له قُدرة مُماثلة على التعبير. وهذا يتضمَّن

الشَّبكات البايزيَّة التي هي النظير الاحتمالي لمنطق القضايا، والشبكات العصبونية، والتي تُعدُّ أساس نهج «التعلُّم المُتعمق» الخاص بالذكاء الاصطناعي.

(٢) المنطق الإسنادي

السؤال التالي هو: هل يُمكننا إنشاء لغة منطقية ذات قدرة أكبر على التعبير؟ إننا نُريد واحدةً من المُمكن فيها إخبار النظام المُعتمد على المعرفة بقواعد لعبة جو على النحو التالي:

«لكل» المواضع على اللوح، و«لكل» الخطوات الزمنية، ها هي القواعد …

إن المنطق الإسنادي، الذي قدَّمه عالم الرياضيات الألماني جوتلوب فريجه في عام ١٨٧٩، يُتيح للمرء كتابة القواعد بهذه الطريقة.[6] إن الاختلاف الأساسي بين منطق القضايا والمنطق الإسنادي هو الآتي: في حين أنَّ النوع الأول يفترض أن العالم يتكون من قضايا صحيحة أو خاطئة، يفترض النوع الثاني أن العالم مُكوَّن من «عناصر» يُمكن «ربطها» معًا بطرقٍ مُتنوِّعة. على سبيل المثال، من المُمكن أن تكون هناك مواضعُ مُجاورة لبعضها، وأوقات تلي بعضها على نحوٍ مُتتالٍ، وقطعُ لعبٍ في مواضع في أوقات معيَّنة، وحركات مسموح بها في أوقات معينة. يسمح المنطق الإسنادي للمرء بالتأكيد على أن خاصية ما صحيحة بالنسبة «لكل» العناصر في العالم؛ ومن ثم يُمكن للمرء كتابة الآتي:

لكلِّ الخطوات الزمنية «ز»، ولكلِّ المواضع «م»، وللونين «ل»، إذا كان دور «ل» في اللعب في الوقت «ز» والموضع «م» خاليًا في الوقت «ز»، فإنه من المسموح به بالنسبة لـ «ل» لعب قطعة لعبٍ في الموضع «م» في الوقت «ز».

مع بعض المحاذير الإضافية وبعض الجُمل الأُخرى التي تعرف مواضع لوح اللعب واللونين ومعنى كلمة «خالٍ»، تكون لدينا بدايات القواعد الكاملة للعبة جو. وستجد أنَّ القواعد المكتوبة باستخدام المنطق الإسنادي ستشغل تقريبًا نفس المساحة التي تشغلها عند كتابتها باللغة الإنجليزية.

إن تطوير «البرمجة المنطقية» في أواخر سبعينيات القرن الماضي وفَّر تقنيةً رائعةً وفعَّالة للتفكير المنطقي والتي تجسَّدت في لغةٍ برمجيةٍ تُسمَّى «برولوج». عرف علماء الكمبيوتر كيف يجعلُون التفكير المنطقي في تلك اللغة يعمل بمُعدَّل ملايين خطوات التفكير في الثانية، مما جعل العديد من التطبيقات المنطقية عملية. وفي عام ١٩٨٢، أعلنت

الحكومة اليابانية عن استثمارٍ هائل في مشروع خاص بالذكاء الاصطناعي قائم على تلك اللغة يُسمى «مشروع الجيل الخامس»،[7] وردَّت الولايات المتحدة الأمريكية والمملكة المتحدة بجهود مُشابِهة.[8،9]

لسوء الحظ، فقد مشروع الجيل الخامس والمشروعات المشابهة زخمه، في أواخر ثمانينيات وأوائل تسعينيات القرن الماضي، جزئيًّا بسبب عدم قدرة المنطق على التعامل مع معلومات غير مؤكَّدة. ولقد جسَّدت تلك المشروعات مُصطلحًا سرعان ما عُد انتقاصيًّا؛ وهو مُصطلح «الذكاء الاصطناعي الجيد القديم الطراز».[10] وشاع اعتبار المنطق غير ذي صلة بالذكاء الاصطناعي؛ في واقع الأمر، لا يعرف العديد من باحثي الذكاء الاصطناعي العاملين الآن في مجال التعلُّم المتعمِّق أيَّ شيء عن المنطق. وهذا الشيوع يبدو أنه مُرشَّح للاختفاء؛ فإذا قبلت بأن العالم به عناصر مُرتبطة ببعضها بطرق متنوعة، فإن المنطق الإسناديَّ سيُصبح ذا صلة، لأنه يوفر الجوانب الرياضية الأساسية للعناصر والعلاقات. وهذا الرأي هو ما يعتقدُه ديمس هاسابس، المدير التنفيذي لشركة ديب مايند التابعة لشركة جوجل:[11]

> يُمكنك النظر إلى التعلُّم المتعمق بالحال الذي هو عليه اليوم باعتباره المكافئ في الدماغ للقشرتين الدماغيتَين الحسيتَين الخاصين بنا؛ القشرة الدماغية البصرية والقشرة الدماغية السمعية. لكن، بالطبع، الذكاء الحقيقي أكثر من ذلك بكثير، فعلينا إعادة جمعه مع التفكير الرمزي والتفكير الأعلى مُستوى، وهي أشياء عديدة حاول الذكاء الاصطناعي الكلاسيكي التعامل معها في ثمانينيات القرن الماضي.
>
> نريد [لتلك النظم] الاستعداد التدريجي لهذا المُستوى الرمزي من التفكير؛ الرياضيات واللغة والمنطق. ومن ثمَّ فهذا جزء كبير من عملنا.

ومن ثم فأحد الدروس المستفادة المهمة من أول ثلاثين عامًا من البحث في مجال الذكاء الاصطناعي هو أنَّ أيَّ برنامج يعرف أشياء، بأيِّ نحوٍ مُفيد، سيحتاج قُدرة على التمثيل والتفكير يُمكن على الأقل مُقارنتها بتلك التي يُتيحها المنطق الإسنادي. وحتى الآن، نحن لا نعرف الشكل الدقيق الذي سيتَّخذُه ذلك؛ إنه يُمكن دمجُه في نُظم تفكيرٍ احتمالي أو نُظُم تعلُّم مُتعمِّق أو تصميمٍ ما هجين لم يظهر للنُّور بعد.

الملحق «ج»: عدم اليقين والاحتمال

في حين أنَّ المنطق يُوفِّر أساسًا عامًّا للتفكير فيما يتعلَّق بمعرفةٍ مُحدَّدة؛ فإنَّ نظرية الاحتمال تتضمَّن التفكير فيما يتعلَّق بمعلوماتٍ غير مُؤكَّدة (والتي تُعدُّ المعرفة المحدَّدة حالة خاصَّة منها). إن عدم اليقين يُعدُّ الموقف المعرفي الطبيعي لأيِّ كيانٍ في العالم الواقعي. وعلى الرغم من أنَّ الأفكار الأساسية للاحتمال جرى تطويرها في القرن السابع عشر، فقط مؤخَّرًا أصبح من المُمكن تمثيل نماذج احتمال كبيرة على نحوٍ صوريٍّ والتفكير فيه.

(١) أسس الاحتمال

تشترك نظريةُ الاحتمال مع المنطق في فكرة أنَّ هناك عوالم مُمكنة. عادةً ما يبدأ المرء بتعريف ماهيتها؛ على سبيل المثال، إن كنتُ أقذف بحجر نردٍ عاديٍّ سداسي الأوجُه، فهناك ستة عوالم (والتي تُسمَّى في بعض الأحيان «نواتج»): ١ و٢ و٣ و٤ و٥ و٦. سيكون واحد منها على وجه التحديد صحيحًا، لكنَّني لا أعرف أيها على نحو مسبق. تفترض نظرية الاحتمال أنه من الممكن إعطاء احتمالٍ لكلِّ عالم؛ بالنِّسبة إلى مثال قذف حجر النرد، سأُعطي احتمالًا قدره ١ / ٦ لكلِّ عالم. (تصادف هنا أن تلك الاحتمالات مُتساوية، لكن ليس من المُفترض أن تكون هكذا في كل الأحوال؛ المُتطلب الوحيد هو أن يُساوي حاصل جمع الاحتمالات ١.) والآن، يُمكنني طرح سؤال مثل: «ما احتمال ظهور عددٍ زوجيٍّ؟» للإجابة على هذا، سأجمع ببساطة احتمالات العوالم الثلاثة التي يكون فيها العدد زوجيًّا؛ وذلك كما يلي: ١ / ٦ + ١ / ٦ + ١ / ٦ = ١ / ٢.

من المنطقيِّ أيضًا أن يجري أخذُ أيِّ أدلةٍ جديدة في الاعتبار. افترض أن عرافًا أخبرني بأن نتيجة قذف النرد ستكون عددًا أوليًّا (أي، ٢ أو ٣ أو ٥). وهذا يستبعدُ العوالم

١ و٤ و٦. إنني ببساطة سآخُذ الاحتمالات المُرتبطة بالعوالم الممكنة المتبقية وأزيد وزن كلٍّ منها بحيث يظلُّ حاصلُ الجمع الإجمالي ١. والآن احتمال كلٍّ من ٢ و٣ و٥ سيساوي ١ / ٣، واحتمال أن يكون ناتج عملية القذف عددًا زوجيًّا ١ / ٣ فقط؛ حيث إن ٢ هو العدد الزوجي الوحيد المتبقِّي في هذه الحالة. إن تلك العملية الخاصة بتحديث الاحتمالات مع ظهور أدلةٍ جديدة تُعدُّ مثالًا على التحديث البايزي.

ومن ثمَّ فهذه الأفكار الخاصَّة بالاحتمال تبدو بسيطةً للغاية! وحتى أي كمبيوتر يُمكنُه جمع الأعداد، إذن، أين المُشكلة؟ تظهر المشكلة عندما يكون هناك أكثر من بضعة عوالم. على سبيل المثال، إن قذفت النرد مائة مرة، فسيكون هناك ٦^١٠٠ ناتج. إنه لأمر غير عمليٍّ بدءُ عملية التفكير الاحتمالي بإعطاء رقم لكلٍّ من هذه النواتج على نحوٍ فردي. ويأتي مفتاح التعامُل مع هذا التعقيد من حقيقة أنَّ عمليات قذف النرد «مُستقلة» إن لم يكن النردُ مغشوشًا؛ أي إن ناتج أيِّ عملية قذف واحدة لن يُؤثِّر على احتمالات نواتج أي عملية قذف أخرى. ومن ثمَّ فالاستقلال مُفيد في إعطاء احتمالات لمجموعات معقدة من الأحداث.

افترض أنني ألعب لعبة مونوبولي مع ابني جورج. إن قطعتي تقف على مربع «مجرد زيارة» وجورج يمتلك المجموعة الصفراء التي عقاراتها على بُعد ١٦ و١٧ و١٩ مربعًا من مُربعي. هل عليه شراء منازل للمجموعة الصفراء الآن، حتى يكون عليَّ أن أدفع له إيجارًا كبيرًا إن وقفت على تلك المربعات، أم عليه الانتظار حتى الدور القادم؟ هذا يعتمد على احتمال الوقوف على المجموعة الصفراء في دوري الحالي.

فيما يلي قواعد قذف النرد في هذه اللعبة: يجري قذف حجري نرد وتتحرك قطعة اللعب وفقًا لإجمالي العددَين الظاهرين؛ إن كان الزوج مُتطابقًا، يقذفهما اللاعب مرة أخرى ويتحرك ثانية؛ وإن تكرر نفس الأمر في المرة الثانية، يقذف اللاعب الحجرين للمرة الثالثة ويتحرَّك ثانية (لكن إن تكرر الأمر في المرة الثالثة، يذهب اللاعب إلى السجن). ومن ثم، على سبيل المثال، قد أحصُل على ٤–٤ ثم ٥–٤، بإجمالي ١٧؛ أو ٢–٢ ثم ٢–٢ ثم ٦–٢، بإجمالي ١٦. وكما أوضحت قبل ذلك، عليَّ أن أجمع ببساطة احتمالات كل العوالم المُنتمية إلى المجموعة الصفراء. لسوء الحظ، هناك العديد من العوالم. وحيث إنه يُمكن قذف ستَّة أحجار نردٍ معًا؛ فإنَّ العوالم قد تكون في عداد الآلاف. وعلاوة على ذلك، لم تعُد عمليات قذف النرد مستقلةً لأنَّ عملية القذف الثانية لن تحدُث ما لم يكن ناتج حجري النرد مُتشابهًا. وعلى الجانب الآخر، إن ضبطنا قيمتَي الزوج الأول من النرد، فستكون قيمتا الزوج الثاني من النرد مُستقلتَين. هل هناك طريقة لتمثيل هذا النوع من الاعتمادية؟

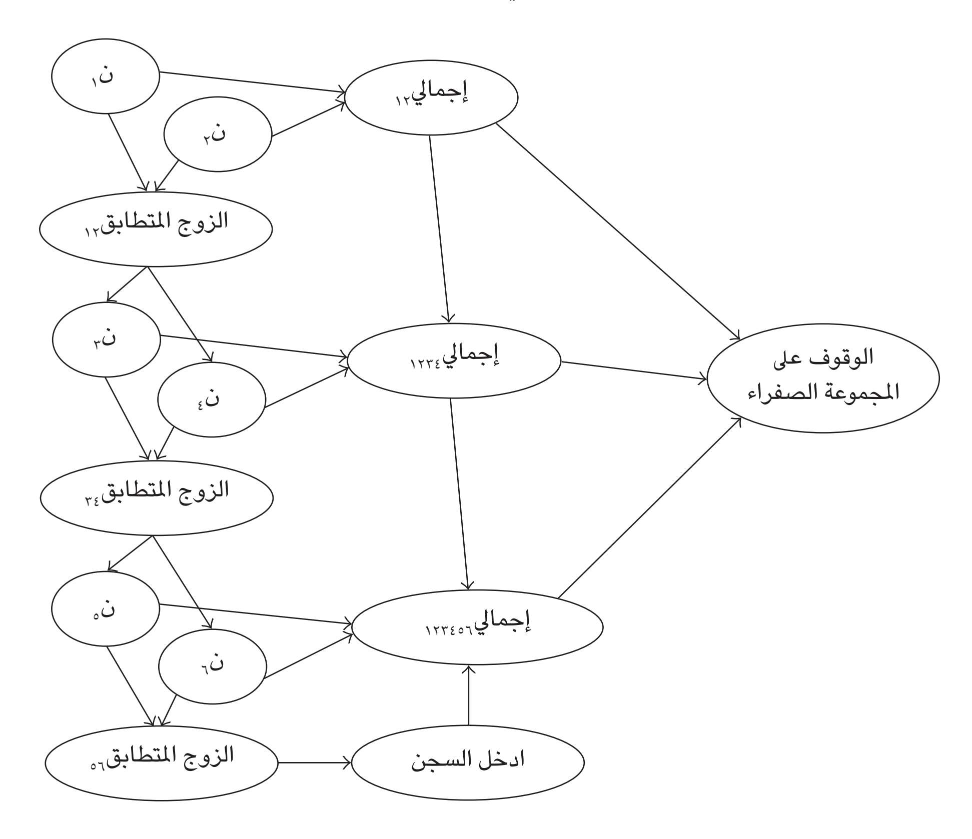
(٢) الشبكات البايزية

في أوائل ثمانينيات القرن الماضي، اقترح جوديا بيرل لغةً صورية سماها «الشبكات البايزية» (التي عادة ما تُختصر إلى شبكات بايز) والتي جعلت من المُمكن، في الكثير من المواقف الواقعية، تمثيل احتمالات عددٍ كبير جدًّا من النواتج على نحو دقيق للغاية.[1]

يعرض الشكل ١ شبكة بايزية تصف قذف النرد في لعبة مونوبولي. إن الاحتمالات الوحيدة التي يجب تحديدها هي احتمالات ١ / ٦ الخاصة بالقيم ١ و٢ و٣ و٤ و٥ و٦ لمرات قذف النرد الفردية (ن١ ون٢ ... إلخ)؛ أي ٣٦ عددًا بدلًا من آلاف الأعداد. إن شرح المعنى الدقيق للشبكة يتطلب معرفة القليل من العمليات الرياضية،[2] لكن الفكرة الأساسية هي أنَّ الأسهُم تُشير إلى علاقات «الاعتمادية»؛ على سبيل المثال، قيمة الزوج المتطابق١٢ تعتمد على قيمتي ن١ ون٢. وبالمثل، تعتمد قيمتا ن٣ ون٤ (مرة القذف التالية لحجري النرد) على الزوج المُتطابق١٢ لأنه إن كان للزوج المتطابق١٢ قيمة «خاطئة»، فستكون قيمة ن٣ ون٤ صفرًا (أي لن تُوجَد مرة قذف تالية).

كما هو الحال مع منطق القضايا، هناك خوارزميات يُمكنها الإجابة عن أي سؤال بالنسبة إلى أي شبكة بايزية بالاستعانة بأيِّ أدلة. على سبيل المثال، يُمكن طلب معرفة احتمال «الوقوف على المجموعة الصفراء»، والذي يتضح أنه يُساوي نحو ٣٫٨٨ بالمائة. (هذا يعني أن جورج يُمكنه الانتظار قبل شراء منازل المجموعة الصفراء.) وعلى نحوٍ طموح أكثر، يُمكننا طلب معرفة احتمال «الوقوف على المجموعة الصفراء» مع الوضع في الاعتبار أن مرة القذف «الثانية» ستكون زوجًا مُتطابقًا يتمثل في العدد ٣. تستنتج الخوارزمية أنه، في هذه الحالة، لا بد أن مرة القذف الأولى نتج عنها زوج متطابق وتخلص إلى أن الإجابة تُساوي ٣٦٫١ بالمائة تقريبًا. هذا مثال على التحديث البايزي؛ عندما يُضاف دليل جديد (والمتمثل هنا في أن مرة القذف الثانية كانت زوجًا متطابقًا متمثلًا في العدد ٣)، يتغير احتمال «الوقوف على المجموعة الصفراء» من ٣٫٨٨ بالمائة إلى ٣٦٫١ بالمائة. بالمثل، يساوي احتمال قذفي للنرد ثلاث مرات (الزوج المتطابق٣٤ صحيح) ٢٫٧٨ بالمائة، في حين يُساوي نفس هذا الاحتمال مع الوضع في الاعتبار الوقوف على المجموعة الصفراء ٢٠٫٤٤ بالمائة.

تُوفِّر الشبكات البايزية سبيلًا لإنشاء نُظمٍ قائمة على المعرفة تتجنَّب أوجُه القُصور التي كانت موجودة في النظم الخبيرة القائمة على القواعد التي ظهرت في ثمانينيات القرن الماضي. (في واقع الأمر، لو قلَّت مقاومة مجتمع الذكاء الاصطناعي للاحتمال في أوائل ثمانينيات القرن الماضي، لتجنُّب فترة التراجع التي تعرَّض لها مجال الذكاء الاصطناعي



شكل ١: شبكة بايزية تمثل قواعد قذف النرد في لعبة مونوبولي وتتيح لخوارزمية حساب احتمال الوقوف على مجموعة معيَّنة من المربعات (مثل المجموعة الصفراء) انطلاقًا من مربع ما آخر (مثل «مجرد زيارة»). (من أجل التبسيط، حذفت الشبكة احتمال الوقوف على مربع «حظ» أو «صندوق الجماعة» والتحول إلى مكان آخر.) يُمثل ن١ ون٢ مرة القذف الأولى لحجري النَّرد، وهما مُستقلَّان (أي لا يُوجد رابط بينهما). إن كان الزوج مُتطابقًا (الزوج المتطابق١٢)، فسيُلقي اللاعب النرد مرة أخرى، ومن ثمَّ تكون قيمتا ن٣ ون٤ غير صفرية، وهكذا. في الوضع الموصوف، يقف اللاعب على المجموعة الصفراء إن كان أي من القيم الإجمالية الثلاثة ١٦ أو ١٧ أو ١٩.

والتي تلت فقاعة النُّظم الخبيرة القائمة على القواعد.) لقد ظهرت آلاف التطبيقات، في مجالاتٍ تتراوح بين التشخيص الطبي ومنع الإرهاب.[3]

توفر الشبكات البايزية آليات لتمثيل الاحتمالات الضرورية وإجراء العمليات الحسابية المطلوبة لتنفيذ التحديث البايزي للعديد من المهام المعقَّدة. لكن كما هو الحال بالنسبة إلى منطق القضايا، إنها محدودة إلى حدٍّ ما في قدرتها على تمثيل المعرفة العامة. في العديد من التطبيقات، يُصبح تمثيل الشبكة البايزية كبيرًا وتكراريًّا للغاية؛ على سبيل المثال، تمامًا كما أن قواعد لعبة جو يجب تكرارها لكل مربَّع في منطق القضايا، يجب تكرار قواعد لعبة مونوبولي القائمة على الاحتمال لكلِّ لاعب ولكل موضع قد يقف عليه أيُّ لاعب ولكل حركة في اللعبة. وتلك الشبكات الهائلة مُستحيل تقريبًا إنشاؤها يدويًّا؛ بدلًا من ذلك، سيكون على المرء اللجوء إلى شفرة مكتوبة بلغة تقليدية مثل «سي+++» لإنتاج مقاطع بايزية متعدِّدة وجمعها معًا. وفي حين أن هذا أمر عملي باعتباره حلًّا هندسيًّا لمشكلة معينة، فإنه يعدُّ عقبةً أمام العمومية؛ لأن شفرة تلك اللغة تجب كتابتها مرةً أخرى على يد خبيرٍ بشريٍّ لكلِّ تطبيق.

(٣) اللغات الاحتمالية القائمة على المنطق الإسنادي

اتضح، لحُسن الحظ، أننا يُمكننا دمج قدرة المنطق الإسنادي على التعبير مع قدرة الشبكات البايزية على تمثيل المعلومات الاحتمالية على نحوٍ دقيق. وهذا المزيج يوفر لنا أفضل ما في العالَمَين؛ النُّظُم «الاحتمالية» القائمة على المعرفة تستطيع التعامُل مع نطاقٍ أكبر بكثير من المواقف الواقعية من أيٍّ من الأساليب المنطقية أو الشبكات البايزية. على سبيل المثال، يُمكننا بسهولة تمثيل معرفة احتمالية مُتعلِّقة بالوراثة كما يلي:

لكل الأفراد «ج» و«ب» و«م»،
إذا كان «ب» أبا «ج»، وكانت «م» أم «ج»،
وكانت فصيلة دم كل من «ب» و«م» AB،
فإن «ج» ستكون فصيلة دمه AB باحتمال ٠٫٥.

إن هذا المزج بين المنطق الإسنادي والاحتمال يُعطينا حقًّا أكثر من مجرَّد طريقةٍ للتعبير عن معلومات غير مؤكَّدة عن العديد من العناصر. إن السبب يكمُن في أننا عندما نضيف عدم يقينٍ إلى عوالم تشتمل على عناصر، فإننا نحصُل على نوعَين جديدَين من عدم اليقين؛ ليس فقط عدم اليقين بشأن ما إذا كانت الحقائق صحيحةً أم خاطئة، وإنَّما أيضًا عدم اليقين بشأن أيِّ العناصر موجودة وعدم اليقين بشأن هُوية كلٍّ منها. وهذان النوعان

من عدم اليقين شائعان بشدة. فالعالم لم يظهر وبه قائمة بالشخصيات، مثل المسرحية الفيكتورية؛ بدلًا من ذلك، إنك تعلم تدريجيًّا بوجود العناصر من خلال الملاحظة.

في بعض الأحيان، يُمكن أن تكون المعرفة الخاصة بالعناصر الجديدة محدَّدة بعض الشيء، مثل عندما تفتح نافذة فندقك وترى كنيسة القلب المقدَّس لأول مرة؛ أو ربما تكون غير مُحددة تمامًا، مثل عندما تشعر بهزة بسيطة والتي قد تكون بسبب زلزال أو قطار مترو مارٍّ. وفي حين أن هُوية الكنيسة واضحة إلى حدٍّ ما، فإن هوية قطارات المترو ليست كذلك؛ فقد تركب نفس القطار الفعلي مئات المرات دون أن تُدرك على الإطلاق أنه نفس القطار. في بعض الأحيان، نحن لا نكون بحاجة إلى تبديد عدم اليقين: أنا عادة لا أُحدِّد أسماء كلِّ الطماطم الموجودة في كيس من طماطم الكرز ولا أتتبع حال كلٍّ منها، إلا إذا كنتُ على الأرجح أسجل تقدُّم تجربة تعفُّن خاصة بالطماطم. أما بالنسبة إلى قاعة ممتلئة بطلاب الدراسات العليا، على الجانب الآخر، فأنا أسعى بقوة إلى تتبُّع هوياتهم. (في إحدى المرات، كان هناك مُساعدان بحثيان في مجموعتي لهما نفس الاسم الأول والاسم الأخير، وكان مظهرهما مُتشابهًا جدًّا، ويعملان على موضوعات مُرتبط بعضها ببعض بشدة؛ على الأقل، كنت متأكدًا بعض الشيء من أنهما كانا شخصَين.) تكمن المشكلة في أننا نُدرك على نحوٍ مباشر ليس «هوية» العناصر، ولكن (جوانب من) «مظهرها»؛ إن العناصر لا تمتلك في الغالب لوحات ترخيص صغيرة تُحدِّد هُويتها على نحوٍ مُميَّز. إن الهوية هي شيء أحيانًا تنسبه عقولنا إلى العناصر من أجل أغراضنا الخاصة.

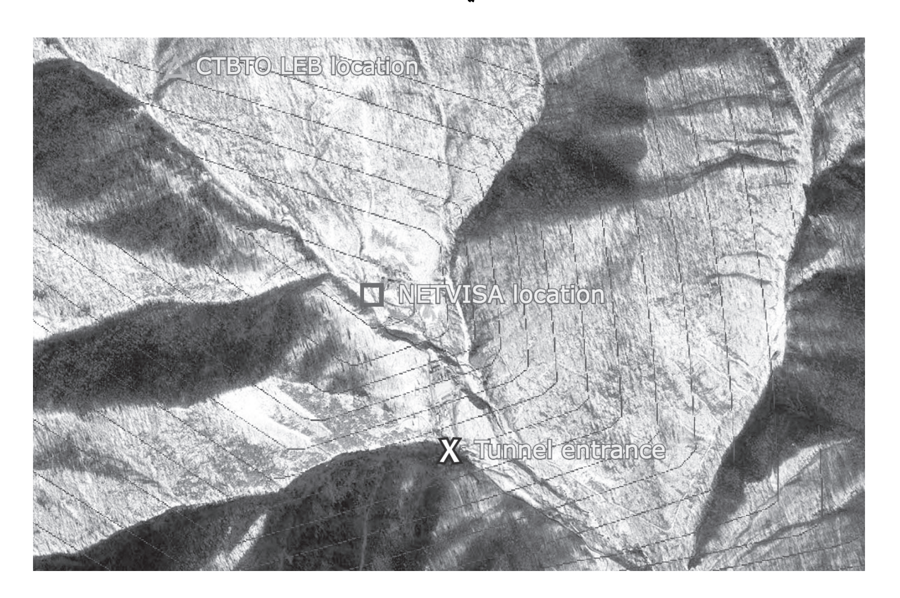
إن المزج بين نظرية الاحتمال ولغة صورية تعبيرية يُعدُّ مجالًا فرعيًّا جديدًا بعض الشيء من الذكاء الاصطناعي، والذي يُطلق عليه عادةً «البرمجة الاحتمالية».[4] لقد جرى تطوير عشرات عديدة من اللغات البرمجية الاحتمالية، والتي يستمدُّ الكثيرُ منها قدرته التعبيرية من اللغات البرمجية العادية وليس من المنطق الإسنادي. إن كل النُّظُم القائمة على اللغات البرمجية الاحتمالية لديها القُدرة على تمثيل المعرفة المعقَّدة غير المؤكَّدة والتفكير فيها. تتضمَّن التطبيقات نظام «ترو سيكل» الخاص بشركة مايكروسوفت، الذي يُقيِّم ملايين لاعبي ألعاب الفيديو كل يوم؛ ونماذج لجوانب المعرفة البشرية التي لم يكن لها تفسير في السابق باستخدام أيِّ فرضية آلية مثل القدرة على تعلُّم فئات عناصر بصرية جديدة من أمثلة فردية؛[5] والمراقبة العالمية للأحداث الزلزالية من أجل مُعاهدة الحظر الشامل للتجارب النووية، وهي المعاهدة المسئولة عن اكتشاف التفجيرات النووية الخفية.[6]

يجمع نظام المراقبة التابع لمُعاهدة الحظر الشامل للتجارب النووية بياناتٍ لحظية خاصة بحركة الأرض عبر شبكة عالمية تتكوَّن من أكثر من ١٥٠ مقياس زلازل ويهدف لاكتشاف كل الأحداث الزلزالية التي تحدث على كوكب الأرض والتي تزيد قوتها عن حدٍّ مُعيَّن وتحديد المشبوه منها. من الواضح أنَّ هناك الكثير من عدم اليقين الخاص بالوجود في هذه المُشكلة؛ لأنَّنا لا نعرف مقدمًا الأحداث التي ستقع؛ علاوة على ذلك، الغالبية العُظمى من الإشارات في البيانات تكون مجرَّد ضوضاء. وهناك أيضًا الكثير من حالات عدم اليقين الخاص بالهوية؛ إن إشارة خاصة بالطاقة الزلزالية المرصُودة في المحطة «أ» الموجودة في القارة القُطبية الجنوبية قد تأتي أو لا تأتي من نفس الحدث الذي جاءت منه الإشارة الأخرى المرصودة في المحطة «ب» الموجودة في البرازيل. إن رصد حركة الأرض يُشبه رصد آلاف المُحادثات الآنية التي حدث خلط بينها بسبب الأصداء والتأخيرات الخاصَّة بالنقل وطغت عليها أصوات الأمواج المتلاطمة.

كيف نحلُّ هذه المشكلة باستخدام البرمجة الاحتمالية؟ قد يعتقد المرء أننا بحاجة إلى بعض الخوارزميات الذكية جدًّا لترتيب كل الاحتمالات. في واقع الأمر، باتِّباع نهج النُّظُم القائمة على المعرفة، لا يكون علينا ابتكار أيِّ خوارزميات جديدة على الإطلاق. إننا ببساطة نستخدم لغةً برمجية احتمالية للتعبير عما نعرفُه عن الجيوفيزياء؛ معدل تكرار حدوث الأحداث في مناطق النشاط الزلزالي الطبيعي ومدى سرعة انتقال الموجات الزلزالية عبر الأرض ومدى سرعة اختفائها ومدى حساسية أدوات الاكتشاف ومدى الضوضاء الموجودة. وبعد ذلك، نُضيف البيانات ونشغل خوارزمية تفكير احتمالي. ونظام المراقبة الناتج، المُسمَّى «نت–فيزا»، كان يعمل باعتباره جزءًا من نظام التحقق من تطبيق المعاهدة منذ عام ٢٠١٨. ويعرض الشكل ٢ اكتشاف نظام «نت–فيزا» لتجربة نووية حدثت في عام ٢٠١٣ في كوريا الشمالية.

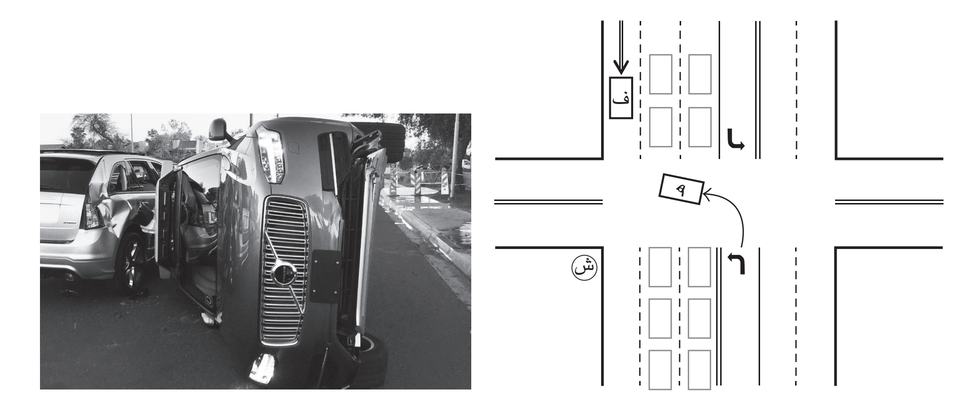
(٤) تتبُّع العالم

يتمثَّل أحد أهم أدوار التفكير الاحتمالي في تتبُّع أجزاء العالم التي تكون غير قابلة للملاحظة على نحوٍ مُباشر. في أغلب ألعاب الفيديو والألعاب اللَّوحية، هذا غير ضروري؛ لأنَّ كلَّ المعلومات ذات الصلة تكون قابلةً للمُلاحظة، لكن في العالم الواقعي نادرًا ما يكون هذا هو الحال.



شكل ٢: تقديرات الموقع الخاصة بالتجربة النووية التي حدثت في ١٢ فبراير من عام ٢٠١٣، والتي قامت بها حكومة كوريا الشمالية. جرى رصد مدخل النفق (حرف الإكس الأسود الموجود في الجزء الأوسط السفلي) في صور الأقمار الصناعية. إن تقدير نظام «نت–فيزا» للموقع هو ٧٠٠ متر تقريبًا من مدخل النفق وهو يعتمد بالأساس على إشارات في محطَّاتٍ على بُعد من ٤ إلى ١٠ آلاف كيلومتر. إن الموقع المحدد من قبل LEB الخاص بمعاهدة الحظر الشامل للتجارب النووية هو التقدير المُجمَع عليه من قبل علماء الجيوفيزياء الخبراء.

المثال على ذلك يأتي من إحدى أولى الحوادث الخطيرة التي تتضمَّن سيارةً ذاتية القيادة. لقد وقعت تلك الحادثة جنوب شارع ماكلينتوك في طريق إيست دون كارلوس في مدينة تيمبي بولاية أريزونا في الرابع والعشرين من مارس عام ٢٠١٧.[7] كما هو موضَّح في الشكل ٣، سيارة ذاتية القيادة من طراز فولفو (ف)، متجهة جنوبًا في شارع ماكلينتوك، اقتربت من تقاطُع تحوَّل فيه للتوِّ لون الإشارة المرورية إلى اللون الأصفر. حارة السيارة الفولفو كانت خالية، لذا، فقد تقدمت بنفس السرعة عبر التقاطع. ثم ظهرت سيارة غير مرئية حاليًّا — السيارة التي من طراز هوندا (هـ) — من خلف صف المرور المتوقف وحدث التصادم.



شكل ٣: (على اليمين) مخطط للوضع الذي أدَّى إلى وقوع الحادث. لقد كانت السيارة الفولفو الذاتية القيادة (ف)، تقترب من أحد التقاطعات، وتسير في الحارة الموجودة في أقصى اليمين بسرعة ٣٨ ميلًا في الساعة. كانت حركة السير متوقِّفة في الحارتَين الأُخريين وتحوَّل لون الإشارة المرورية (ش) إلى اللون الأصفر. قامت سيارة هوندا (هـ)، والتي لم تكن مرئية للسيارة الفولفو، بانعطافٍ إلى اليسار؛ (على اليسار) نتائج الحادث.

لاستنتاج الوجود المُحتمل لسيارة هوندا غير المرئية، يُمكن للسيارة فولفو تجميع الأدلة عند اقترابها من التقاطع. على وجه الخصوص، المرور في الحارتين الأُخريين مُتوقِّف حتى رغم أنَّ الإشارة خضراء؛ السيارات الموجودة في مقدمة الصف لا تتقدَّم إلى الأمام باتجاه التقاطُع ومصابيح الكبح خاصَّتها مُضاءة. هذا ليس دليلًا «قاطعًا» على وجود سيارة غير مرئية تنعطف إلى اليسار، ولكنَّه لا يجب أن يكون كذلك؛ فحتى الاحتمال القليل يكون كافيًا لاقتراح الإبطاء ودخول التقاطُع على نحوٍ أكثر حذرًا.

إنَّ الغاية من هذه القصة هي أنَّ الكيانات الذكية العاملة في بيئات قابلة للمُلاحظة على نحوٍ جزئيٍّ يجبُ أن تحتسب لما لا يُمكنها رؤيته — قدر ما يُمكنها — اعتمادًا على الأدلة المُستمدة مما يُمكنها رؤيته.

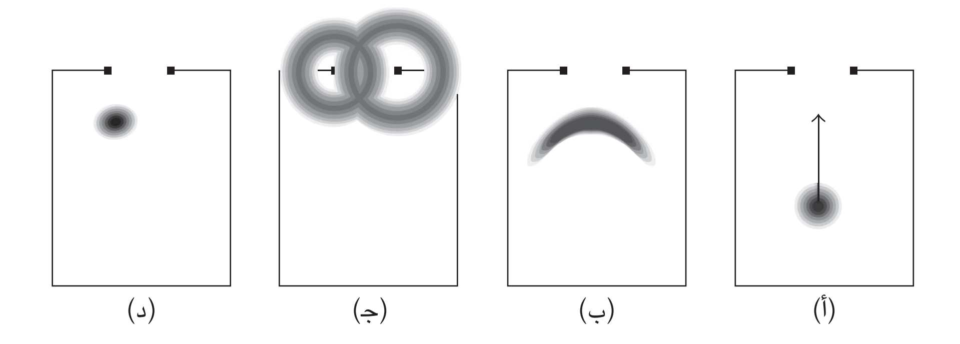
إليك مثال آخر أقرب إليك: أين تُوجَد مفاتيحك؟ ما لم يتصادف قيادتُك لسيارتك أثناء قراءة هذا الكتاب — وهو الأمر غير المُحبَّذ — فأنت على الأرجح لا يُمكنُك رؤيتها الآن. على الجانب الآخر، أنت على الأرجح تعرف مكانها؛ إنها في جيبك أو حقيبتك أو على الطاولة المُجاورة للسرير أو في جيب معطفك المعلق أو ربما على المشجب في المطبخ.

أنت تعرف هذا لأنك وضعتها هناك ولم يتغيَّر مكانها منذ ذلك الوقت. هذا مثال بسيط لاستخدام المعرفة والتفكير لتتبُّع حالة العالم.

بدون هذه القدرة، سنشعُر بالضياع؛ غالبًا حرفيًّا تمامًا. على سبيل المثال، في وقت كتابتي لهذه السطور، أنا أنظر إلى الحائط الأبيض لغرفة في فندق لا ملامح له. أين أنا؟ إن كان عليَّ الاعتماد على مدخلاتي الإدراكية الحالية، فسأشعُر بالضياع بالفعل. في حقيقة الأمر، أنا أعرف أنني في زيورخ لأنَّني وصلتُ إليها أمس ولم أترُكها. إن الروبوتات، شأنها شأن البشر، يجب أن تعرف أين هي حتى يُمكنها إيجاد طريقها بنجاح عبر الغرف والمباني والشوارع والغابات والصحاري.

في الذكاء الاصطناعي، نحن نستخدم مُصطلح «الحالة المعرفية» للإشارة إلى معرفة الكيان الحالي لحالة العالم؛ بصرف النظر عن درجة عدم الاكتمال وعدم اليقين التي هي عليها. بوجهٍ عام، الحالة المعرفية — وليس المُدخلات الإدراكية الحالية — هي الأساس الصحيح لصنع القرارات فيما يتعلَّق بما علينا فعله. إن تحديث تلك الحالة نشاط حيوي لأيِّ كيانٍ ذكي. وبالنِّسبة إلى بعض أجزاء تلك الحالة، يحدُث هذا تلقائيًّا؛ على سبيل المثال، بدا لي للتوِّ أنَّني في زيورخ، دون أن يكون عليَّ التفكير في الأمر. بالنسبة إلى أجزاء أخرى، يحدُث التحديث عند الطلب، إن جاز التعبير. على سبيل المثال، عندما أستيقظ في مدينة جديدة وأعاني من تعبٍ شديدٍ بسبب اختلاف التوقيت، في منتصف رحلة طويلة، قد يكون عليَّ القيام بجهدٍ واعٍ لإدراك أين أنا، وما أنا بصدد القيام به، ولماذا؛ وهذا، على ما أعتقد، يُشبه بعض الشيء قيام الكمبيوتر المحمول بإعادة تشغيل نفسه. إن التتبُّع لا يعني المعرفة «الدقيقة» الدائمة لحالة «كل شيء» في العالم. من الواضح أن هذا مُستحيل؛ على سبيل المثال، أنا ليست لديَّ أي فكرة عمن يشغل الغرف الأخرى في فندقي الغريب في زيورخ، فضلًا عن المواقع والأنشطة الحالية للجانب الأكبر من الثمانية مليارات شخص الذين يعيشون على كوكب الأرض. وأنا أيضًا ليست لديَّ أدنى فكرة عما يحدُث في باقي الكون فيما يتجاوَّز المجموعة الشمسية. إن عدم يقيني فيما يتعلَّق بالحالة الحالية للأشياء هائل وحتمي.

إن الطريقة الأساسية لتتبُّع عالم غير مؤكَّد هي «التحديث البايزي». عادةً ما تُنفِّذ الخوارزميات التي تقوم بهذا خطوتَين؛ خطوة خاصة بالتوقُّع، يتوقَّع فيها الكيان الحالة الحالية للعالم في ضوء أحدث تحركاته، ثم خطوة خاصة بالتحديث، حيث يستقبل مدخلاتٍ إدراكية جديدة ويحدِّث مُعتقداته تبعًا لذلك. لتوضيح كيف يعمل هذا، تأمَّل



شكل ٤: روبوت يُحاول السير من مُنتصف الغرفة والخروج من الباب. (أ) الحالة المعرفية الأولية: الروبوت غير مُتيقِّن على نحوٍ ما من موقعه؛ إنه يُحاول التحرُّك مترًا ونصف باتجاه الباب. (ب) الخطوة الخاصة بالتوقع: يُقدِّر الروبوت أنه قريب من الباب ولكنه غير مُتيقن تمامًا من الاتجاه الذي سار فيه بالفعل؛ لأن محركاته قديمة وعجلاته غير مستقرة. (ج) يقيس الروبوت المسافة لكل من عضادتي الباب باستخدام جهاز سونار جودته ضعيفة؛ التقديرات هي ٧٠ سنتيمترًا من عضادة الباب اليسرى و٨٥ سنتيمترًا من العضادة اليمنى. (د) الخطوة الخاصة بالتحديث: إن الجمع بين التوقع في الشكل (ب) والملاحظة التي في الشكل (ج) يعطينا الحالة المعرفية الجديدة. والآن، الروبوت لديه فكرة جيدة جدًّا عن المكان الموجود فيه وسيحتاج إلى تصحيح مساره قليلًا للخروج عبر الباب.

معي المشكلة التي يُواجهها أيُّ روبوت فيما يتعلَّق بتحديد المكان الموجود فيه. يوضح الشكل ٤(أ) مثالًا نموذجيًّا لهذا الأمر: الروبوت موجود في مُنتصف إحدى الغرف، ولديه بعض عدم اليقين فيما يتعلَّق بموقعه الدقيق، ويُريد الخروج عبر الباب. إنه يأمر عجلاته بالتحرُّك لمسافة مترٍ ونصف باتجاه الباب؛ لسوء الحظ، عجلاته قديمة وغير مُستقرَّة، لذا توقُّع الروبوت بشأن المكان الذي سينتهي إليه غير مؤكد تمامًا، كما هو موضَّح في الشكل ٤(ب). إن حاول التحرُّك الآن، فقد يصطدم بشيء. لحُسن الحظ، لديه جهاز سونار لقياس المسافة إلى عضادتي الباب. كما يُوضِّح الشكل ٤(ج)، تقترح القياسات أن الروبوت يُوجَد على بُعد نحو ٧٠ سنتيمترًا من عضادة الباب اليُسرى و٨٥ سنتيمترًا من العضادة اليمنى. وفي النهاية، يُحدِّث الروبوت حالته المعرفية بالجمع بين التوقُّع في الشكل ٤(ب) والقياسات الموجودة في الشكل ٤(ج) للحصول على الحالة المعرفية الجديدة البادية في الشكل ٤(د).

إن خوارزمية تتبع الحالة المعرفية يُمكن تطبيقها لمعالجة ليس فقط عدم اليقين بشأن الموقع، وإنما أيضًا عدم اليقين بشأن الخريطة نفسها. ينتج عن هذا أسلوب يُسمَّى «تحديد الموقع وبناء الخريطة في آنٍ واحد». إن هذا الأسلوب مكوِّن رئيسي للعديد من تطبيقات الذكاء الاصطناعي، التي تتراوح بين نظم الواقع المُعزَّز والسيارات الذاتية القيادة وعربات الاستكشاف الكوكبية.

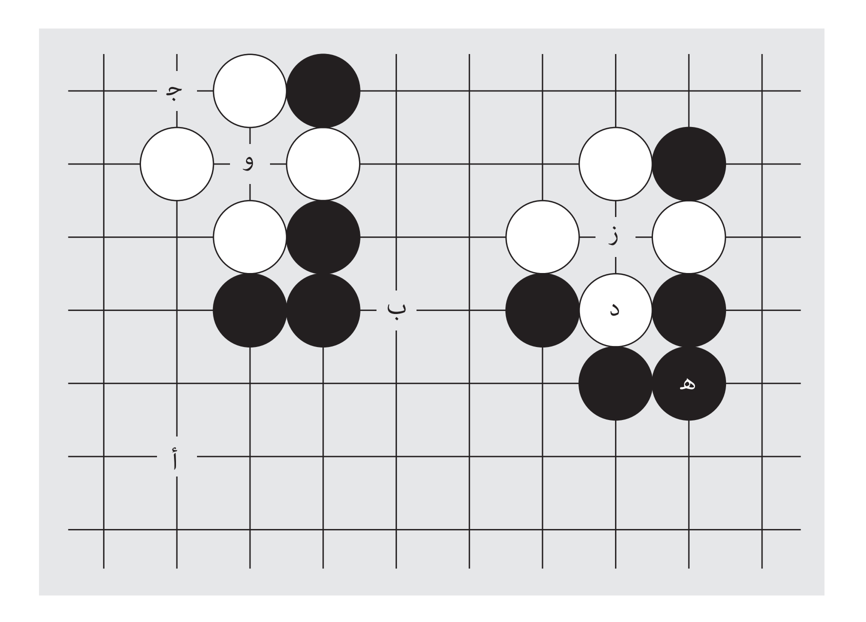
الملحق «د»: التعلم من التجربة

يعني التعلُّم تحسين الأداء بناءً على التجربة. بالنسبة إلى نظام إدراك بصري، قد يعني هذا تعلُّم تمييز المزيد من فئات العناصر اعتمادًا على رؤية أمثلة لتلك الفئات؛ بالنسبة إلى نظامٍ قائم على المعرفة، يُعدُّ مجرَّد اكتساب المزيد من المعرفة شكلًا من التعلُّم؛ لأنه يعني أن النظام يمكنه الإجابة عن المزيد من الأسئلة؛ بالنسبة إلى نظام اتخاذ قرارٍ استباقي مثل «ألفا جو»، يُمكن أن يعني التعلُّم تحسين قدرته على تقييم الأوضاع أو تحسين قدرته على استكشاف أجزاء مُفيدة من شجرة الاحتمالات.

(١) التعلم من الأمثلة

يُسمَّى أكثر أشكال تعلُّم الآلة شيوعًا التعلم «الموجَّه». تُعطى أيُّ خوارزمية قائمة على التعلم الموجه مجموعة من الأمثلة التدريبية، والتي تُسمى كلٌّ منها حسب الناتج الصحيح، ويجب أن تنتج فرضية تتعلَّق بماهية القاعدة الصحيحة. عادةً، يسعى أيُّ نظامٍ قائم على التعلُّم الموجَّه إلى تحسين التوافُق بين الفرضية والأمثلة التدريبية. وفي الغالب، تكون هناك أيضًا عقوبة على الفرضيات المُعقَّدة أكثر مما هو ضروري؛ كما هو مُوصى به من قبل مبدأ القصد.

دعني أُعطِ مثالًا على ذلك فيما يتعلَّق بمُشكلة تعلُّم الحركات المسموح بها في لعبة جو. (إن كنت تعرف بالفعل قواعد تلك العبة، فسيكون على الأقل تتبُّع ما هو معروض



شكل ١: الحركات المسموح وغير المسموح بها في لعبة جو؛ الانتقال إلى المواضع «أ» و«ب» و«ج» مسموح به بالنسبة إلى اللاعب الأسود، في حين أنَّ الانتقال للمواضع «د» و«ه» و«و» غير مسموح به. الانتقال للموضع «ز» قد يكون أو لا يكون مسموحًا به، اعتمادًا على ما جرى في السابق في اللعبة.

هنا سهلًا؛ وإن لم يكن الأمر كذلك، فستكون قادرًا أكثر على التعاطف مع برنامج التعلم.) افترض أنَّ الخوارزمية تبدأ بالفرضية الآتية:

> لكل الخطوات الزمنية «ز»، ولكل المواضع «م»، من المسموح به وضع قطعة لعب في الموضع «م» في الوقت «ز».

إنه دور اللاعب الأسود للانتقال إلى الوضع الموضَّح في الشكل ١. تجرب الخوارزمية الموضع «أ»؛ هذا جيد. والموضعان «ب» و«ج» جيدان أيضًا. ثم تجرب الموضع «د»، وهو موضع تُوجَد عليه قطعة لعب بيضاء؛ هذا غير مسموح به. (في لعبتي الشطرنج والطاولة، سيكون هذا لا بأس به؛ فهذه هي الطريقة التي يجري بها الاستحواذ على القطع.) إن الانتقال إلى الموضع «ه»، وهو الموضع الذي تُوجَد به قطعة لعب سوداء، غير مسموح به

أيضًا. (إنه غير مسموح به في الشطرنج أيضًا، لكن مسموح به في لعبة الطاولة.) والآن، من خلال تلك الأمثلة التدريبية الخمسة، قد تقترح الخوارزمية الفرضية التالية:

> لكل الخطوات الزمنية «ز»، ولكل المواضع «م»، إذا كان «م» خاليًا في الوقت «ز»، فإنه من المسموح به وضع قطعة لعب في الموضع «م» في الوقت «ز».

وبعد ذلك، تجرب الموضع «و» وتندهش عندما تجد أن الانتقال إليه غير مسموح به. وبعد بعض البدايات الخاطئة، تستقرُّ على ما يلي:

> لكل الخطوات الزمنية «ز»، ولكل المواضع «م»، إذا كان «م» خاليًا في الوقت «ز» وكان «م» غير مُحاط بقطع لعب خاصة بالمنافس، فإنه من المسموح به وضع قطعة لعب في الموضع «م» في الوقت «ز».

(تُسمَّى هذه أحيانًا بقاعدة «عدم الانتحار».) وفي النهاية، تُجرِّب الموضع «ز»، والذي يتضح أنه مسموح بالانتقال إليه. وبعد التفكير لبعض الوقت وربما القيام بالقليل من التجارب الأخرى، تستقرُّ على الفرضية التي ترى أن الموضع «ز» جيد، حتى وإن كان مُحاطًا بقطع لعب المنافس؛ لأنه يؤدي إلى الاستحواذ على قطعة اللعب البيضاء الموجودة في الموضع «د»؛ ومن ثمَّ يصبح غير محاطٍ بأي قطع لعب للمنافس على الفور.

كما يمكن أن تلاحظ من خلال التطور التدريجي للقواعد، تحدث عملية التعلم من خلال سلسلة من التعديلات التي تتمُّ على الفرضية حتى تتوافق مع الأمثلة الملحوظة. هذا شيء تستطيع أي خوارزمية تعلم فعله بسهولة. لقد صمَّم الباحثون في مجال تعلم الآلة كل أشكال الخوارزميات المبتكرة لإيجاد فرضياتٍ جيدة بسرعة. هنا، الخوارزمية تبحث في مجال التعبيرات المنطقية التي تُمثِّل قواعد لعبة جو، لكن الفرضيات يُمكنها أيضًا أن تُكوِّن تعبيراتٍ جبريةً تُمثِّل قوانين فيزيائية أو شبكات بايزية احتمالية تُمثِّل الأمراض والأعراض أو حتى برامج كمبيوتر تُمثِّل السلوك المُعقد لآلةٍ أخرى.

هناك نقطة ثانية مُهمَّة تتمثَّل في أنه «حتى الفرضيات الجيدة يمكن أن تكون خاطئة»؛ في واقع الأمر، الفرضية المذكورة سلفًا خاطئة، حتى بعد تعديلها لضمان أن الانتقال إلى الموضع «ز» حركة مسموح بها. إنها يجب أن تتضمَّن قاعدة «الأو» أو «عدم التكرار»؛ على سبيل المثال، إن كان اللاعب الأبيض قد استحوذ للتوِّ على قطعة لعب سوداء عند الموضع «ز» بالانتقال للموضع «د»، فقد لا يُعيد اللاعب الأسود الاستحواذ بالانتقال

إلى الموضع «ز» حيث إن هذا يُنتج نفس الوضع ثانية. لاحظ أن تلك القاعدة تُعدُّ انحرافًا جذريًّا عما تعلَّمه البرنامج حتى الآن؛ لأنَّ هذا يعني أن ما هو مسموح به لا يُمكن تحديده من الوضع الحالي؛ بدلًا من ذلك، يجب على المرء أيضًا تذكُّر الأوضاع السابقة.

أشار الفيلسوف الاسكتلندي ديفيد هيوم في عام ١٧٤٨ إلى أن الاستقراء — أي التفكير الذي من خلاله يُمكن الوصول من ملاحظات محدَّدة إلى مبادئ عامة — لا يمكن أبدًا ضمان صحَّته.[1] في النظرية الحديثة للتعلم الإحصائي، نحن لا نطلب ضمانات للصحة التامة، وإنما فقط ضمانًا بأن الفرضية التي جرى التوصل إليها «على الأرجح صحيحة على نحو تقريبي».[2] يمكن لخوارزمية التعلُّم أن تكون «غير محظوظة» وترى عينة غير مُمثلة؛ على سبيل المثال، قد لا تجرب أبدًا حركة مثل الانتقال إلى الموضع «ز»، مُعتقدةً أن تلك الحركة غير مسموح بها. وقد تفشل أيضًا في توقع بعض الحالات المتطرفة الغريبة، مثل تلك المتضمنة في بعض الأشكال الأكثر تعقيدًا والنادر ظهورها من قاعدة عدم التكرار.[3] لكن ما دام الكون يُوفِّر درجةً ما من الانتظام، فمن غير المحتمل جدًّا أن تنتج الخوارزمية فرضية سيئة للغاية؛ لأنَّ مثل هذه الفرضية كانت على نحو مُرجَّح جدًّا «ستُكتشف» من قبل إحدى التجارب.

يُعدُّ التعلُّم المُتعمِّق — وهو التقنية التي تسبَّبت في كل هذه الضجة التي أُثيرت عن الذكاء الاصطناعي في وسائل الإعلام — بالأساس شكلًا من أشكال التعلُّم الموجه. إنه يُمثِّل أحد أهم النجاحات التي تحقَّقت في مجال الذكاء الاصطناعي في العقود الأخيرة، لذا من المُهم فهمُ كيف يعمل. علاوة على ذلك، يعتقد بعض الباحثين أنه سيُؤدِّي إلى إنتاج نُظم ذكاءٍ اصطناعي مُضاهية للذكاء البشري في خلال بضعة أعوام، لذا، من المُهم تقييم ما إذا كان من المُحتمل أن يكون هذا صحيحًا أم لا.

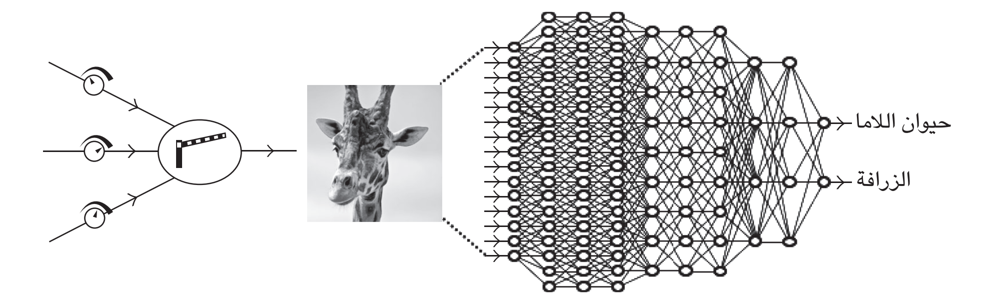
من الأسهل فهمُ التعلُّم المُتعمق في سياق مُهمَّة مُعيَّنة؛ على سبيل المثال، تعلم كيفية التمييز بين الزراف وحيوانات اللاما. ففي ضوء وجود بعض الصور الفوتوغرافية المُعنونة لكلٍّ منهما، يكون على خوارزمية التعلُّم إنشاء فرضية تسمح لها بتصنيف الصور غير المُعنونة. إن أيَّ صورة، من وجهة نظر الكمبيوتر، ليست سوى جدولٍ كبيرٍ من الأعداد، كلُّ عدد منها يُماثل إحدى قِيَم الآر جي بي الثلاث لكلِّ بكسل من الصورة. لذا، بدلًا من وجود فرضية خاصة بلعبة جو تأخذ أحد أوضاع اللوح وإحدى الحركات كمدخلات وتُقرِّر ما إذا كانت الحركة مسموحًا بها أم لا، نحتاج إلى فرضية خاصَّة بالزراف وحيوانات اللاما تأخُذ جدولًا من الأعداد كمُدخلاتٍ وتتنبَّأ بفئة «الزراف أو حيوانات اللاما».

السؤال الآن هو: أيُّ نوعية من الفرضيات تلك التي نحتاجُها؟ على مدى الخمسين عامًا الأخيرة أو نحو ذلك من البحث في مجال الرؤية الحاسوبية، جرت تجربة العديد من الأساليب. الأسلوب السائد حاليًّا هو «الشبكة الالتفافية المتعمقة». دعني أوضح لك ما يعنيه هذا؛ إنها تُسمى «شبكة» لأنها تُمثِّل تعبيرًا رياضيًّا مُعقَّدًا مُؤلَّفًا بطريقة منتظمة من العديد من التعبيرات الفرعية الأصغر، والهيكل التركيبي له شكل الشبكة. (عادة ما يطلق على تلك الشبكات «الشبكات العصبونية» لأنَّ مُصمِّميها يستمدُّون إلهامهم من شبكات العصبونات الموجودة في الدماغ.) وهي تُوصف بأنها «التفافية» لأن هذه طريقة رياضية مُنمقة للقول بأنَّ هيكل الشبكة يُكرِّر نفسه بنمطٍ ثابت عبر صورة المُدخلات بالكامل. وتُوصف بأنها «مُتعمِّقة»؛ لأنَّ تلك الشبكات تشتمل في الغالب على عدة طبقات، ولأنها تبدُو رائعة ومُخيفة قليلًا.

يظهر مثال مُبسَّط في الشكل ٢؛ إنه مُبسَّط لأنَّ الشبكات الحقيقية قد تكون لها مئات الطبقات وملايين التفرُّعات. إن الشبكة في واقع الأمر عبارة عن صورة لتعبيرٍ رياضي مُعقَّد وقابل للتعديل. كل تفرُّع في الشبكة يقابل تعبيرًا بسيطًا قابلًا للتَّعديل، كما هو موضَّح في الشكل. تجري التعديلات بتغيير «الأوزان» في كلِّ مدخل، كما هو مُحدَّد من قبل «عناصر التحكُّم في الحجم». ثم يجري تمرير المجموع المرجَّح للمُدخلات عبر دالة مرور قبل الوصول لجانب المُخرجات الخاص بالتفرُّع؛ في الغالب، تتجاوز دالة المرور القيم الصَّغيرة وتسمح فقط بالقيم الأكبر.

يحدُث التعلُّم في الشبكة ببساطة بتعديل كلِّ أزرار عناصر التحكُّم في الحجم لتقليل خطأ التنبؤ في الأمثلة المعنونة. إن الأمر بسيط للغاية؛ لا تُوجَد أيُّ حِيَل ولا خوارزميات بارعة على نحوٍ خاصٍّ. إن تحديد الاتجاه الذي ستُدار فيه الأزرار لتقليل الخطأ لهو تطبيق بسيط لقواعد التفاضُل والتكامل لحساب كيف سيؤدِّي تغيير كل وزن إلى تغيير الخطأ في طبقة المُخرجات. وهذا يُؤدي إلى صيغة بسيطة لنقل الخطأ إلى الخلف من طبقة المُخرجات إلى طبقة المدخلات، مع ضبط الأزرار في أثناء ذلك.

على نحوٍ إعجازي، تنجح العملية. وبالنسبة إلى مهمة تمييز العناصر الموجودة في الصور، أبدت خوارزميات التعلم المتعمق أداءً رائعًا. ظهرت أولى بوادر هذا في تحدي إيمدج نت لعام ٢٠١٢ الذي وفَّر بيانات تدريبية تتكون من ١٫٢ مليون صورة مُعنونة من ألف فئة ثم تطلُب من الخوارزمية عنونة مائة ألف صورة جديدة.[4] كان جيوف هينتون، وهو عالم نفسٍ حوسبي بريطاني، من طليعة المشاركين في أول ثورة في مجال الشبكات



شكل ٢: (على اليمين) تصوير مُبسَّط لشبكة التفافية مُتعمِّقة خاصة بتمييز العناصر في الصور. تجري تغذية قيم بكسلات الصور من اليسار وتُنتج الشبكة القيم عند التفرُّعين الموجودين في أقصى اليمين، مما يُشير إلى مدى احتمال أن تكون الصورة حيوان لاما أو زرافة. لاحظ كيف أن نمط الروابط الداخلية، المشار إليه بالخطوط السوداء في الطبقة الأولى، يتكرَّر عبر الطبقة بأكملها (على اليسار)؛ هذا هو أحد تفرُّعات الشبكة. هناك وزن قابل للتعديل لكلِّ قيمةٍ مُدخلة، الذي يُحدِّد للتفرُّع قدر الانتباه الذي يجب أن يُوليه لها. وبعد ذلك، تمرُّ الإشارة المدخلة الإجمالية عبر دالة مرور تسمح بمرور الإشارات الكبيرة خلالها، ولكن تتجاوز الإشارات الصغيرة.

العصبونية في ثمانينيات القرن العشرين، مع شبكةٍ الْتفافية مُتعمِّقة كبيرة للغاية؛ إذ كانت تتكوَّن من ٦٥٠ ألف تفرُّع و٦٠ مليون مُعامِل. وصل هو ومجموعته في جامعة تورونتو إلى معدَّل خطأ إيمدج نت يصل إلى ١٥ بالمائة، وهو ما يُعدُّ تطوُّرًا هائلًا في ضوء أفضل معدَّل سابق جرى الوصول إليه والذي تمثَّل في ٢٦ بالمائة.[5] وبحلول عام ٢٠١٥، كانت عشرات الفرق تستخدم طرق التعلُّم المُتعمِّق، وقد قلَّ مُعدَّل الخطأ إلى ٥ بالمائة، والذي يُشبِّه ذلك الخاص بالمبحوث الذي قضى أسابيع في تعلُّم كيفية التمييز بين الألف فئة في الاختبار.[6] وبحلول عام ٢٠١٧، كان معدَّل خطأ الآلة ٢ بالمائة.

تقريبًا في نفس هذه الفترة، حدثت تطوُّرات مشابهة في تمييز الكلام والترجمة الآلية باستخدام طرق مُماثلة. وإن جمعنا هذه المجالات الثلاثة معًا، فسنجدُ أنها من أهم المجالات التطبيقية في عالم الذكاء الاصطناعي. وقد لعب التعلُّم المُتعمق أيضًا دورًا مُهمًّا في تطبيقات التعلُّم المُعزَّز؛ على سبيل المثال، في تعلُّم دالة التقييم التي يستخدمها «ألفا جو» لتقدير مدى مرغوبية الأوضاع المُستقبلية المُمكنة، وفي تعلُّم أدوات التحكم في سلوكيات الروبوتات المُعقَّدة.

حتى هذه اللحظة، نحن لدَينا فهم قليل للغاية للسبب وراء عمل التعلم المتعمق على النحو الجيد الذي هو عليه. ربما يتمثَّل أفضل تفسير في أنَّ الشبكات المتعمقة عميقة؛ فنظرًا لأنها تتكوَّن من طبقات مُتعدِّدة، فيمكن لكلِّ طبقة أن تتعلَّم تحولًا بسيطًا نسبيًّا من مُدخلاتها إلى مخرجاتها، في حين تتجمَّع تلك التحوُّلات البسيطة المتعدِّدة لتُشكِّل التحوُّل المعقَّد المطلوب للانتقال من صورةٍ ما إلى اسم فئة. بالإضافة إلى ذلك، الشبكات المتعمقة الخاصَّة بالرؤية لديها هيكل داخلي يفرض الثبات الانتقالي والثبات الحجمي؛ بمعنى أن الكلب كلب بصرف النَّظر عن مكان ظهوره في الصورة وبصرف النظر عن الحجم الذي يبدو به فيها.

هناك خاصية مُهمَّة أخرى للشبكات المُتعمقة والمتمثِّلة في أنها عادةً ما يبدو أنها تكتشف تمثيلاتٍ داخلية تُجسِّد السمات الأساسية للصور مثل العيون والخطوط والأشكال البسيطة. لا تكون أيٌّ من تلك السِّمات مُضمنة. نحن نعرف أنها موجودة لأننا بإمكاننا العمل مع الشبكة المُدربة ومعرفة أنواع البيانات التي تجعل التفرعات الداخلية (عادة تلك التي تكون قريبة من طبقة المُخرجات) حيوية. في الحقيقة، من المُمكن تشغيل خوارزمية التعلم بطريقةٍ مُختلفة بحيث تعدل الصورة نفسها لإنتاج ردٍّ أقوى في تفرُّعات داخلية مختارة. إن تكرار تلك العملية عدة مرات ينتج ما هو معروف الآن بصور «الاستهلال» (تيمنًا بفيلم «استهلال» (انسبشن) أو «الحلم العميق»)، مثل تلك التي تظهر في الشكل ٣.[7] لقد أصبح الاستهلال شكلًا فنيًّا في حدِّ ذاته، والذي يُنتِج صورًا تختلف تمامًا عن الأشكال الفنية البشرية الأخرى.

رغم كل الإنجازات الملحوظة لنُظُم التعلُّم المتعمق، فإنها، بحسب فهمنا لها حاليًّا، بعيدة كل البُعد عن توفير أساس للنُّظُم الذكية العامة. إن نقطة الضعف الأساسية فيها تتمثَّل في أنها عبارة عن «دوائر»؛ فهي تعدُّ نظراء لمنطق القضايا والشبكات البايزية، التي، رغم كل خصائصها الرائعة، تفتقد القُدرة على التعبير عن أشكال معقَّدة من المعرفة على نحوٍ دقيق. هذا يعني أنَّ الشبكات المتعمقة العاملة في «الوضع الأصلي» تتطلَّب كمياتٍ هائلة من الدوائر لتمثيل أنواع بسيطة نسبيًّا من المعرفة العامة. وهذا، بدوره، يعني ضمنيًّا ضرورة تعلُّم أعدادٍ هائلة من الأوزان؛ ومن ثمَّ الحاجة لعددٍ غير معقول من الأمثلة؛ أكثر مما يُمكن أن يُوفِّره الكون.

يرى البعض أن الدماغ يتكوَّن أيضًا من دوائر، عناصرها هي العصبونات؛ ومن ثمَّ يُمكن أن تدعم الدوائر الذكاء المُضاهي للذكاء البشري. هذا صحيح، ولكن فقط في



شكل ٣: صورة مُنتجة من قبل برنامج «ديب دريم» الخاص بشركة جوجل.

نفس الإطار الذي يرى أن الأدمغة مصنوعة من ذرات؛ يُمكن للذرات في واقع الأمر دعم الذكاء المضاهي للذكاء البشريِّ، لكن هذا لا يعني أن مجرَّد تجميع العديد من الذرات معًا سينتج ذكاءً. فيجب ترتيب الذرَّات بطرقٍ مُعيَّنة. وعلى نفس النحو، يجب ترتيب الدوائر بطرقٍ مُعيَّنة. وأجهزة الكمبيوتر مصنوعة أيضًا من دوائر، فيما يتعلَّق بذاكراتها ووحدات المعالجة الخاصة بها، لكن تلك الدوائر يجب ترتيبها بطرقٍ مُعيَّنة، وتجب إضافة طبقات من البرمجيات، قبل أن يكون بإمكان الكمبيوتر دعم تشغيل نُظُم التفكير المنطقي واللغات البرمجية العالية المُستوى. ولكن، في الوقت الحالي، لا يُوجد ما يُشير إلى أن نُظُم التعلُّم المُتعمق يُمكنها تطوير تلك القدرات بنفسها؛ كما أنه لا معنى من الناحية العلمية لأن نطلُب منها فعل ذلك.

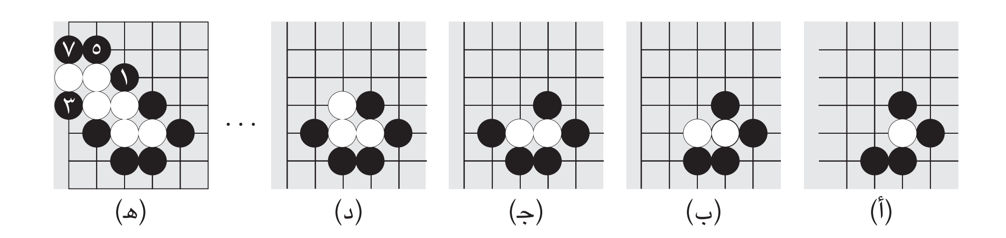
هناك أسباب أخرى للاعتقاد بأن التعلم المتعمق قد يصل لُمستوًى ثابتٍ ما بعيدٍ كل البُعد عن الذكاء العام، لكن لا يتَّسع المقام هنا لتشخيص كل المشكلات؛ فلقد ذكر آخرون، داخل[8] وخارج[9] مجال التعلُّم المتعمق، الكثير منها. الفكرة هي أن مجرَّد إنشاء شبكات أكبر وأعمق ومجموعات بيانات وآلات أكبر ليس كافيًا لإيجاد ذكاءٍ اصطناعي مُضاهٍ للذكاء البشري. لقد رأينا بالفعل (في الملحق «ب») وجهة نظر ديمس هاسابس، المدير التنفيذي لشركة ديب مايند، التي ترى أن «التفكير الرمزي والتفكير الأعلى مُستوى» أساسيان بالنسبة إلى الذكاء الاصطناعي. وهناك خبيرُ تعلُّم مُتعمق بارز آخر يُدعى فرانسوا شوليه صاغ الأمر على النحو التالي:[10] «هناك الكثير من التطبيقات البعيدة المنال تمامًا بالنِّسبة إلى أساليب التعلُّم المُتعمق الحالية؛ حتى في وجود كمياتٍ هائلة من البيانات المفسَّرة من قبل البشر. ... نحن بحاجة للابتعاد عن تخطيطات المُدخلات إلى المخرجات البسيطة والاتجاه إلى التفكير والتجريد».

(٢) التعلُّم من التفكير

عندما يلحُّ عليك التفكير في شيءٍ ما، فهذا يرجع إلى أنك لا تعرف بالفعل الإجابة. فعندما يسألك شخصٌ ما عن رقم هاتفك المحمول الجديد تمامًا، فأنت على الأرجح لن تعرفه. وستقول في نفسك: «حسنًا، أنا لا أعرفه؛ ومن ثم، كيف سأجده؟» وحيث إنك لستَ مرتبطًا بشدة بالهاتف المحمول، فأنت لا تعرف كيف تجده. وستقول في نفسك: «كيف يُمكنني معرفة كيفية إيجاده؟» ستكون لديك إجابة عامة على هذا السؤال: «إنهم على الأرجح يضعونه في مكان ما يسهل على المُستخدمين إيجاده». (بالطبع، قد تكون مخطئًا بهذا الشأن.) الأماكن الأكثر احتمالاً ستتمثَّل في الجزء العلوي من الشاشة الرئيسية (إنه غير موجود هناك) أو داخل تطبيق الهاتف أو في قسم «الإعدادات» الموجود في هذا التطبيق. ستُجرِّب الانتقال إلى قسم «الإعدادات» ثم إلى قسم «الهاتف»، وستجده هناك.

في المرة التالية التي ستُسأل فيها عن رقم هاتفك، إما ستكون على علمٍ به وإما ستعرف على وجه التحديد كيف ستجدُه. إنك ستتذكر طريقة إيجاده، ليس فقط بالنسبة «لهذا» الهاتف في «هذا» الموقف، ولكن أيضًا «لكل» الهواتف المُماثلة في «كل» المواقف؛ أي ستخزن وتُعيد استخدام حلٍّ «عامٍّ» للمُشكلة. إن هذا التعميم مُبرَّر لأنك أدركتَ أن تفاصيل هذا الهاتف بعينه وهذا الموقف بعينه غير ذات صلة. وستُصدَم إن نجحت الطريقة التي تبنَّيتها فقط في أيام الثلاثاء بالنسبة لأرقام الهواتف المنتهية بالرقمَين ١٧.

توفر لعبة جو مثالًا جميلًا على هذا النوع من التعلم. في الشكل ٤(أ)، نرى موقفًا شائعًا حيث يُهدِّد اللاعب الأسود بالاستحواذ على قطعة لعب اللاعب الأبيض بالإحاطة بها. يُحاول اللاعب الأبيض الهروب بإضافة قطع لعب قريبة من قطعة اللعب الأصلية، لكن اللاعب الأسود يستمرُّ في قطع الطرق المُؤدية للهُروب. يُشكِّل هذا النمط من الحركات «سلمًا» من قطع اللعب على نحوٍ قُطري عبر اللوح، حتى يصل إلى الحافة؛ وحينها، لا يجد اللاعب الأبيض أي موضع ينتقل إليه. إن كنت اللاعب الأبيض، فأنت على الأرجح لن ترتكب نفس الخطأ مرة أخرى؛ ستُدرك أن نمط السلم «دائمًا» ما تنتج عنه في النهاية عملية استحواذ؛ وذلك بالنِّسبة «إلى أي» وضع أوَّلي و«أي» اتجاه، وفي «أي» مرحلة من اللعبة، سواء كنتَ أنت اللاعب الأبيض أو الأسود. الاستثناء الوحيد يحدُث عندما يؤدي السُّلَّم إلى بعض قطع اللعب الإضافية التي تنتمي إلى الشخص الهارب. وتنبع عمومية نمط السُّلَّم على نحوٍ مباشر من قواعد لعبة جو.



شكل ٤: مفهوم «السُّلم» في لعبة جو. (أ) يُهدِّد اللاعب الأسود بالاستحواذ على قطعة اللعب الخاصَّة باللاعب الأبيض. (ب) يُحاول اللاعب الأبيض الهروب. (ج) يسد اللاعب الأسود اتجاه الهُروب. (د) يُجرِّب اللاعب الأبيض الاتجاه الآخر. (هـ) يستمر اللعب بالتسلسل المشار إليه بالأرقام. ويصلُ السُّلَّم في النهاية إلى حافة اللوح، حيث لا يُوجَد موضع يُمكن أن ينتقل إليه اللاعب الأبيض. الضربة القاضية تمَّت من خلال الحركة رقم ٧؛ مجموعة اللاعب الأبيض جرت الإحاطة بها بالكامل وماتت.

إن مثال رقم الهاتف غير المعروف ومثال السُّلَّم الخاص بلعبة جو يُوضِّحان إمكانية تعلم قواعد عامَّة وفعَّالة من مثالٍ واحد؛ وهو أمر مُختلف تمامًا عن ملايين الأمثلة المطلوبة للتعلُّم المتعمق. في مجال الذكاء الاصطناعي، يطلق على هذا النوع من التعلم

«التعلم القائم على الشرح»؛ فعند رؤية المثال، يستطيع الكيان أن يشرح لنفسه «سبب» حدوثه على النحو الذي هو عليه ويُمكنه استنتاج المبدأ العامِّ بمعرفة العوامل التي كانت أساسية للشرح.

في حقيقة الأمر، هذه العملية لا تُضيف بنفسها معرفة جديدة؛ على سبيل المثال، يستطيع اللاعب الأبيض ببساطة استنتاج وجود وناتج نمط السُّلَّم العام من قواعد لعبة جو، دون أن يرى مُطلقًا مثالًا عليه.[11] لكن الاحتمالات هي أنه لن يكتشف أبدًا مفهوم السُّلَّم دون أن يرى مثالًا عليه؛ ومن ثمَّ، يُمكننا النظر إلى التعلُّم القائم على الشرح باعتباره طريقةً فعَّالة لحفظ نتائج عملية حوسبةٍ بطريقةٍ عامة؛ وذلك من أجل تجنُّب ضرورة إعادة نفس عملية التفكير باختصار (أو ارتكاب نفس الخطأ من خلال عملية تفكير معيبة) في المُستقبل.

لقد أكَّدت الأبحاث في مجال العلوم المعرفية على أهمية هذا النوع من التعلُّم في المعرفة البشرية. فهو يعدُّ، تحت مُسمَّى «التجميع»، أحد الأعمدة الأساسية في نظرية ألن نيويل ذات التأثير الكبير الخاصَّة بالمعرفة.[12] (كان نيويل أحد الحاضرين في ورشة عمل دارتموث التي عقدت في عام ١٩٥٦ وقد فاز بجائزة تورينج لعام ١٩٧٥ بالاشتراك مع هربرت سايمن.) فهو يُفسِّر كيف يُصبح البشر أكثر طلاقةً في المهامِّ المعرفية من خلال الممارسة، حيث إن المهام الفرعية العديدة التي تطلبت في السابق تفكيرًا تُصبح آلية. وبدونه، كانت ستقتصر المحادثات البشرية على ردودٍ مكونة من كلمة أو كلمتَين، وكان الرياضيون سيستمرون في العد على أصابعهم.

ملاحظات

الفصل الأول: ماذا لو نجحنا؟

(1) The first edition of my textbook on AI, co-authored with Peter Norvig, currently director of research at Google: Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, 1st ed. (Prentice Hall, 1995).

(2) Robinson developed the *resolution* algorithm, which can, given enough time, prove any logical consequence of a set of first-order logical assertions. Unlike previous algorithms, it did not require conversion to propositional logic. J. Alan Robinson, "A machine-oriented logic based on the resolution principle," *Journal of the ACM* 12 (1965): 23–41.

(3) Arthur Samuel, an American pioneer of the computer era, did his early work at IBM. The paper describing his work on checkers was the first to use the term *machine learning,* although Alan Turing had already talked about "a machine that can learn from experience" as early as 1947. Arthur Samuel, "Some studies in machine learning using the game of checkers," *IBM Journal of Research and Development* 3 (1959): 210–29.

(4) The "Lighthill Report," as it became known, led to the termination of research funding for AI except at the universities of Edinburgh and Sussex: Michael James Lighthill, "Artificial intelligence: A general survey,"

in *Artificial Intelligence: A Paper Symposium* (Science Research Council of Great Britain, 1973).

(5) The CDC 6600 filled an entire room and cost the equivalent of $20 million. For its era it was incredibly powerful, albeit a million times less powerful than an iPhone.

(6) Following Deep Blue's victory over Kasparov, at least one commentator predicted that it would take one hundred years before the same thing happened in Go: George Johnson, "To test a powerful computer, play an ancient game," *The New York Times*, July 29, 1997.

(7) For a highly readable history of the development of nuclear technology, see Richard Rhodes, *The Making of the Atomic Bomb* (Simon & Schuster, 1987).

(8) A simple supervised learning algorithm may not have this effect, unless it is wrapped within an A/B testing framework (as is common in online marketing settings). Bandit algorithms and reinforcement learning algorithms will have this effect if they operate with an explicit representation of user state or an implicit representation in terms of the history of interactions with the user.

(9) Some have argued that profit-maximizing corporations are already out-of-control artificial entities. See, for example, Charles Stross, "Dude, you broke the future!" (keynote, 34th Chaos Communications Congress, 2017). See also Ted Chiang, "Silicon Valley is turning into its own worst fear," *Buzzfeed*, December 18, 2017. The idea is explored further by Daniel Hillis, "The first machine intelligences," in *Possible Minds: Twenty-Five Ways of Looking at AI*, ed. John Brockman (Penguin Press, 2019).

(10) For its time, Wiener's paper was a rare exception to the prevailing view that all technological progress was a good thing: Norbert Wiener, "Some moral and technical consequences of automation," *Science* 131 (1960): 1355–58.

الفصل الثاني: مفهوم الذكاء في البشر والآلات

(1) Santiago Ramon y Cajal proposed synaptic changes as the site of learning in 1894, but it was not until the late 1960s that this hypothesis was confirmed experimentally. See Timothy Bliss and Terje Lomo, "Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path," *Journal of Physiology* 232 (1973): 331–56.

(2) For a brief introduction, see James Gorman, "Learning how little we know about the brain," *The New York Times*, November 10, 2014. See also Tom Siegfried, "There's a long way to go in understanding the brain," *ScienceNews*, July 25, 2017. A special 2017 issue of the journal *Neuron* (vol. 94, pp. 933–1040) provides a good overview of many different approaches to understanding the brain.

(3) The presence or absence of consciousness — actual subjective experience — certainly makes a difference in our moral consideration for machines. If ever we gain enough understanding to design conscious machines or to detect that we have done so, we would face many important moral issues for which we are largely unprepared.

(4) The following paper was among the first to make a clear connection between reinforcement learning algorithms and neurophysiological recordings: Wolfram Schultz, Peter Dayan, and P. Read Montague, "A neural substrate of prediction and reward," *Science* 275 (1997): 1593–99.

(5) Studies of intracranial stimulation were carried out with the hope of finding cures for various mental illnesses. See, for example, Robert Heath, "Electrical self-stimulation of the brain in man," *American Journal of Psychiatry* 120 (1963): 571–77.

(6) An example of a species that may be facing self-extinction via addiction: Bryson Voirin, "Biology and conservation of the pygmy sloth, *Bradypus pygmaeus*," *Journal of Mammalogy* 96 (2015): 703–7.

(7) The *Baldwin effect* in evolution is usually attributed to the following paper: James Baldwin, “A new factor in evolution,” *American Naturalist* 30 (1896): 441–51.

(8) The core idea of the Baldwin effect also appears in the following work: Conwy Lloyd Morgan, *Habit and Instinct* (Edward Arnold, 1896).

(9) A modern analysis and computer implementation demonstrating the Baldwin effect: Geoffrey Hinton and Steven Nowlan, “How learning can guide evolution,” *Complex Systems* 1 (1987): 495–502.

(10) Further elucidation of the Baldwin effect by a computer model that includes the evolution of the internal reward-signaling circuitry: David Ackley and Michael Littman, “Interactions between learning and evolution,” in *Artificial Life II*, ed. Christopher Langton et al. (Addison-Wesley, 1991).

(11) Here I am pointing to the roots of our present-day concept of intelligence, rather than describing the ancient Greek concept of *nous*, which had a variety of related meanings.

(12) The quotation is taken from Aristotle, *Nicomachean Ethics*, Book III, 3, 1112b.

(13) Cardano, one of the first European mathematicians to consider negative numbers, developed an early mathematical treatment of probability in games. He died in 1576, eighty-seven years before his work appeared in print: Gerolamo Cardano, *Liber de ludo aleae* (Lyons, 1663).

(14) Arnauld’s work, initially published anonymously, is often called *The Port-Royal Logic*: Antoine Arnauld, *La logique, ou l’art de penser* (Chez Charles Savreux, 1662). See also Blaise Pascal, *Pensées* (Chez Guillaume Desprez, 1670).

(15) The concept of utility: Daniel Bernoulli, “Specimen theoriae novae de mensura sortis,” *Proceedings of the St. Petersburg Imperial Academy of*

Sciences 5 (1738): 175–92. Bernoulli's idea of utility arises from considering a merchant, Sempronius, choosing whether to transport a valuable cargo in one ship or to split it between two, assuming that each ship has a 50 percent probability of sinking on the journey. The expected monetary value of the two solutions is the same, but Sempronius clearly prefers the two-ship solution.

(16) By most accounts, von Neumann did not himself invent this architecture but his name was on an early draft of an influential report describing the EDVAC storedprogram computer.

(17) The work of von Neumann and Morgenstern is in many ways the foundation of modern economic theory: John von Neumann and Oskar Morgenstern, *Theory of Games and Economic Behavior* (Princeton University Press, 1944).

(18) The proposal that utility is a sum of discounted rewards was put forward as a mathematically convenient hypothesis by Paul Samuelson, "A note on measurement of utility," *Review of Economic Studies* 4 (1937): 155–61. If s_0, s_1, … is a sequence of states, then its utility in this model is $U(s_0, s_1, \ldots) = \Sigma_t \gamma^t R(s_t)$, where γ is a discount factor and R is a reward function describing the desirability of a state. Naïve application of this model seldom agrees with the judgment of real individuals about the desirability of present and future rewards. For a thorough analysis, see Shane Frederick, George Loewenstein, and Ted O'Donoghue, "Time discounting and time preference: A critical review," *Journal of Economic Literature* 40 (2002): 351–401.

(19) Maurice Allais, a French economist, proposed a decision scenario in which humans appear consistently to violate the von Neumann-Morgenstern axioms: Maurice Allais, "Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'école américaine," *Econometrica* 21 (1953): 503–46.

(20) For an introduction to non-quantitative decision analysis, see Michael Wellman, "Fundamental concepts of qualitative probabilistic networks," *Artificial Intelligence* 44 (1990): 257–303.

(21) I will discuss the evidence for human irrationality further in Chapter 9. The standard references include the following: Allais, "Le comportement"; Daniel Ellsberg, *Risk, Ambiguity, and Decision* (PhD thesis, Harvard University, 1962); Amos Tversky and Daniel Kahneman, "Judgment under uncertainty: Heuristics and biases," *Science* 185 (1974): 1124–31.

(22) It should be clear that this is a thought experiment that cannot be realized in practice. Choices about different futures are never presented in full detail, and humans never have the luxury of minutely examining and savoring those futures before choosing. Instead, one is given only brief summaries, such as "librarian" or "coal miner." In making such a choice, one is really being asked to compare two probability distributions over complete futures, one beginning with the choice "librarian" and the other "coal miner," with each distribution assuming optimal actions on one's own part within each future. Needless to say, this is not easy.

(23) The first mention of a randomized strategy for games appears in Pierre Rémond de Montmort, *Essay d'analyse sur les jeux de hazard,* 2nd ed. (Chez Jacques Quillau, 1713). The book identifies a certain Monsieur de Waldegrave as the source of an optimal randomized solution for the card game Le Her. Details of Waldegrave's identity are revealed by David Bellhouse, "The problem of Waldegrave," *Electronic Journal for History of Probability and Statistics* 3 (2007).

(24) The problem is fully defined by specifying the probability that Alice scores in each of four cases: when she shoots to Bob's right and he dives right or left, and when she shoots to his left and he dives right or left. In this case, these probabilities are 25 percent, 70 percent, 65 percent, and 10 percent respectively. Now suppose that Alice's strategy is to shoot

to Bob's right with probability p and his left with probability $1 - p$, while Bob dives to his right with probability q and left with probability $1 - q$. The payoff to Alice is $U_A = 0.25pq + 0.70\, p(1 - q) + 0.65\,(1 - p)\,q + 0.10\,(1 - p)(1 - q)$, while Bob's payoff is $U_B = -U_A$. At equilibrium, $\partial U_A / \partial p = 0$ and $\partial U_B / \partial q = 0$, giving $p = 0.55$ and $q = 0.60$.

(25) The original game-theoretic problem was introduced by Merrill Flood and Melvin Dresher at the RAND Corporation; Tucker saw the payoff matrix on a visit to their offices and proposed a "story" to go along with it.

(26) Game theorists typically say that Alice and Bob could *cooperate* with each other (refuse to talk) or *defect* and rat on their accomplice. I find this language confusing, because "cooperate with each other" is not a choice that each agent can make separately, and because in common parlance one often talks about cooperating with the police, receiving a lighter sentence in return for cooperating, and so on.

(27) For an interesting trust-based solution to the prisoner's dilemma and other games, see Joshua Letchford, Vincent Conitzer, and Kamal Jain, "An 'ethical' game-theoretic solution concept for two-player perfect-information games," in *Proceedings of the 4th International Workshop on Web and Internet Economics*, ed. Christos Papadimitriou and Shuzhong Zhang (Springer, 2008).

(28) Origin of the tragedy of the commons: William Forster Lloyd, *Two Lectures on the Checks to Population* (Oxford University, 1833).

(29) Modern revival of the topic in the context of global ecology: Garrett Hardin, "The tragedy of the commons," *Science* 162 (1968): 1243–48.

(30) It's quite possible that even if we had tried to build intelligent machines from chemical reactions or biological cells, those assemblages would have turned out to be implementations of Turing machines in nontraditional materials. Whether an object is a generalpurpose computer has nothing to do with what it's made of.

(31) Turing's breakthrough paper defined what is now known as the *Turing machine*, the basis for modern computer science. The *Entscheidungsproblem,* or *decision problem*, in the title is the problem of deciding entailment in first-order logic: Alan Turing, "On computable numbers, with an application to the *Entscheidungsproblem*," *Proceedings of the London Mathematical Society*, 2nd ser., 42 (1936): 230–65.

(32) A good survey of research on negative capacitance by one of its inventors: Sayeef Salahuddin, "Review of negative capacitance transistors," in *International Symposium on VLSI Technology, Systems and Application* (IEEE Press, 2016).

(33) For a much better explanation of quantum computation, see Scott Aaronson, *Quantum Computing since Democritus* (Cambridge University Press, 2013).

(34) The paper that established a clear complexity-theoretic distinction between classical and quantum computation: Ethan Bernstein and Umesh Vazirani, "Quantum complexity theory," *SIAM Journal on Computing* 26 (1997): 1411–73.

(35) The following article by a renowned physicist provides a good introduction to the current state of understanding and technology: John Preskill, "Quantum computing in the NISQ era and beyond," arXiv:1801.00862 (2018).

(36) On the maximum computational ability of a one-kilogram object: Seth Lloyd, "Ultimate physical limits to computation," *Nature* 406 (2000): 1047–54.

(37) For an example of the suggestion that humans may be the pinnacle of physically achievable intelligence, see Kevin Kelly, "The myth of a superhuman AI," *Wired*, April 25, 2017: "We tend to believe that the limit is way beyond us, way 'above' us, as we are 'above' an ant … What evidence do we have that the limit is not us?"

(38) In case you are wondering about a simple trick to solve the halting problem: the obvious method of just running the program to see if it finishes doesn't work, because that method doesn't necessarily finish. You might wait a million years and still not know if the program is really stuck in an infinite loop or just taking its time.

(39) The proof that the halting problem is undecidable is an elegant piece of trickery. The question: Is there a LoopChecker(P, X) program that, for *any* program P and *any* input X, decides correctly, in finite time, whether P applied to input X will halt and produce a result or keep chugging away forever? Suppose that LoopChecker exists. Now write a program Q that calls LoopChecker as a subroutine, with Q itself and X as inputs, and then does the *opposite* of what LoopChecker(Q, X) predicts. So, if LoopChecker says that Q halts, Q doesn't halt, and vice versa. Thus, the assumption that LoopChecker exists leads to a contradiction, so LoopChecker cannot exist.

(40) I say "appear" because, as yet, the claim that the class of NP-complete problems requires superpolynomial time (usually referred to as P ≠ NP) is still an unproven conjecture. After almost fifty years of research, however, nearly all mathematicians and computer scientists are convinced the claim is true.

(41) Lovelace's writings on computation appear mainly in her notes attached to her translation of an Italian engineer's commentary on Babbage's engine: L. F. Menabrea, "Sketch of the Analytical Engine invented by Charles Babbage," trans. Ada, Countess of Lovelace, in *Scientific Memoirs*, vol. III, ed. R. Taylor (R. and J. E. Taylor, 1843). Menabrea's original article, written in French and based on lectures given by Babbage in 1840, appears in *Bibliothèque Universelle de Genève* 82 (1842).

(42) One of the seminal early papers on the possibility of artificial intelligence: Alan Turing, “Computing machinery and intelligence,” *Mind* 59 (1950): 433–60.

(43) The Shakey project at SRI is summarized in a retrospective by one of its leaders: Nils Nilsson, “Shakey the robot,” technical note 323 (SRI International, 1984). A twentyfour-minute film, *SHAKEY: Experimentation in Robot Learning and Planning*, was made in 1969 and garnered national attention.

(44) The book that marked the beginning of modern, probability-based AI: Judea Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (Morgan Kaufmann, 1988).

(45) Technically, chess is not fully observable. A program does need to remember a small amount of information to determine the legality of castling and en passant moves and to define draws by repetition or by the fifty-move rule.

(46) For a complete exposition, see Chapter 2 of Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed. (Pearson, 2010).

(47) The size of the state space for StarCraft is discussed by Santiago Ontañon et al., “A survey of real-time strategy game AI research and competition in StarCraft,” *IEEE Transactions on Computational Intelligence and AI in Games* 5 (2013): 293–311. Vast numbers of moves are possible because a player can move all units simultaneously. The numbers go down as restrictions are imposed on how many units or groups of units can be moved at once.

(48) On human-machine competition in StarCraft: Tom Simonite, “DeepMind beats pros at StarCraft in another triumph for bots,” *Wired*, January 25, 2019.

(49) AlphaZero is described by David Silver et al., “Mastering chess and shogi by self-play with a general reinforcement learning algorithm,” arXiv:1712.01815 (2017).

(50) Optimal paths in graphs are found using the A* algorithm and its many descendants: Peter Hart, Nils Nilsson, and Bertram Raphael, “A formal basis for the heuristic determination of minimum cost paths,” *IEEE Transactions on Systems Science and Cybernetics* SSC-4 (1968): 100–107.

(51) The paper that introduced the Advice Taker program and logic-based knowledge systems: John McCarthy, “Programs with common sense,” in *Proceedings of the Symposium on Mechanisation of Thought Processes* (Her Majesty’s Stationery Office, 1958).

(52) To get some sense of the significance of knowledge-based systems, consider database systems. A database contains concrete, individual facts, such as the location of my keys and the identities of your Facebook friends. Database systems cannot store general rules, such as the rules of chess or the legal definition of British citizenship. They can count how many people called Alice have friends called Bob, but they cannot determine whether a particular Alice meets the conditions for British citizenship or whether a particular sequence of moves on a chessboard will lead to checkmate. Database systems cannot combine two pieces of knowledge to produce a third: they support memory but not reasoning. (It is true that many modern database systems provide a way to add rules and a way to use those rules to derive new facts; to the extent that they do, they are really knowledge-based systems.) Despite being highly constricted versions of knowledge-based systems, database systems underlie most of present-day commercial activity and generate hundreds of billions of dollars in value every year.

(53) The original paper describing the completeness theorem for first-order logic: Kurt Gödel, "Die Vollständigkeit der Axiome des logischen Funktionenkalküls," *Monatshefte für Mathematik* 37 (1930): 349–60.

(54) The reasoning algorithm for first-order logic does have a gap: if there is no answer — that is, if the available knowledge is insufficient to give an answer either way — then the algorithm may never finish. This is unavoidable: it is mathematically *impossible* for a correct algorithm *always* to terminate with "don't know," for essentially the same reason that no algorithm can solve the halting problem (page 37).

(55) The first algorithm for theorem-proving in first-order logic worked by reducing firstorder sentences to (very large numbers of) propositional sentences: Martin Davis and Hilary Putnam, "A computing procedure for quantification theory," *Journal of the ACM* 7 (1960): 201–15. Robinson's resolution algorithm operated directly on first-order logical sentences, using "unification" to match complex expressions containing logical variables: J. Alan Robinson, "A machine-oriented logic based on the resolution principle," *Journal of the ACM* 12 (1965): 23–41.

(56) One might wonder how Shakey the logical robot ever reached any definite conclusions about what to do. The answer is simple: Shakey's knowledge base contained false assertions. For example, Shakey believed that by executing "push object A through door D into room B," object A would end up in room B. This belief was false because Shakey could get stuck in the doorway or miss the doorway altogether or someone might sneakily remove object A from Shakey's grasp. Shakey's plan execution module could detect plan failure and replan accordingly, so Shakey was not, strictly speaking, a purely logical system.

(57) An early commentary on the role of probability in human thinking: Pierre-Simon Laplace, *Essai philosophique sur les probabilités* (Mme. Ve. Courcier, 1814).

(58) Bayesian logic described in a fairly nontechnical way: Stuart Russell, "Unifying logic and probability," *Communications of the ACM* 58 (2015): 88–97. The paper draws heavily on the PhD thesis research of my former student Brian Milch.

(59) The original source for Bayes' theorem: Thomas Bayes and Richard Price, "An essay towards solving a problem in the doctrine of chances," *Philosophical Transactions of the Royal Society of London* 53 (1763): 370–418.

(60) Technically, Samuel's program did not treat winning and losing as absolute rewards; by fixing the value of material to be positive; however, the program generally tended to work towards winning.

(61) The application of reinforcement learning to produce a world-class backgammon program: Gerald Tesauro, "Temporal difference learning and TD-Gammon," *Communications of the ACM* 38 (1995): 58–68.

(62) The DQN system that learns to play a wide variety of video games using deep RL: Volodymyr Mnih et al., "Human-level control through deep reinforcement learning," *Nature* 518 (2015): 529–33.

(63) Bill Gates's remarks on Dota 2 AI: Catherine Clifford, "Bill Gates says gamer bots from Elon Musk-backed nonprofit are 'huge milestone' in A.I.," CNBC, June 28, 2018.

(64) An account of OpenAI Five's victory over the human world champions at Dota 2: Kelsey Piper, "AI triumphs against the world's top pro team in strategy game Dota 2," *Vox*, April 13, 2019.

(65) A compendium of cases in the literature where misspecification of reward functions led to unexpected behavior: Victoria Krakovna, "Specification gaming examples in AI," *Deep Safety* (blog), April 2, 2018.

(66) A case where an evolutionary fitness function defined in terms of maximum velocity led to very unexpected results: Karl Sims, "Evolving

virtual creatures," in *Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques* (ACM, 1994).

(67) For a fascinating exposition of the possibilities of reflex agents, see Valentino Braitenberg, *Vehicles: Experiments in Synthetic Psychology* (MIT Press, 1984).

(68) News article on a fatal accident involving a vehicle in autonomous mode that hit a pedestrian: Devin Coldewey, "Uber in fatal crash detected pedestrian but had emergency braking disabled," *TechCrunch*, May 24, 2018.

(69) On steering control algorithms, see, for example, Jarrod Snider, "Automatic steering methods for autonomous automobile path tracking," technical report CMU-RI-TR-09-08, Robotics Institute, Carnegie Mellon University, 2009.

(70) Norfolk and Norwich terriers are two categories in the ImageNet database. They are notoriously hard to tell apart and were viewed as a single breed until 1964.

(71) A very unfortunate incident with image labeling: Daniel Howley, "Google Photos mislabels 2 black Americans as gorillas," *Yahoo Tech*, June 29, 2015.

(72) Follow-up article on Google and gorillas: Tom Simonite, "When it comes to gorillas, Google Photos remains blind," *Wired*, January 11, 2018.

الفصل الثالث: كيف قد يتطوَّر الذكاء الاصطناعي في المُستقبل؟

(1) The basic plan for game-playing algorithms was laid out by Claude Shannon, "Programming a computer for playing chess," *Philosophical Magazine*, 7th ser., 41 (1950): 256–75.

(2) See figure 5,12 of Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, 1st ed. (Prentice Hall, 1995). Note that

the rating of chess players and chess programs is not an exact science. Kasparov's highest-ever Elo rating was 2851, achieved in 1999, but current chess engines such as Stockfish are rated at 3300 or more.

(3) The earliest reported autonomous vehicle on a public road: Ernst Dickmanns and Alfred Zapp, "Autonomous high speed road vehicle guidance by computer vision," *IFAC Proceedings Volumes* 20 (1987): 221–26.

(4) The safety record for Google (subsequently Waymo) vehicles: "Waymo safety report: On the road to fully self-driving," 2018.

(5) So far there have been at least two driver fatalities and one pedestrian fatality. Some references follow, along with brief quotes describing what happened. Danny Yadron and Dan Tynan, "Tesla driver dies in first fatal crash while using autopilot mode," *Guardian*, June 30, 2016: "The autopilot sensors on the Model S failed to distinguish a white tractor-trailer crossing the highway against a bright sky." Megan Rose Dickey, "Tesla Model X sped up in Autopilot mode seconds before fatal crash, according to NTSB," *TechCrunch*, June 7, 2018: "At 3 seconds prior to the crash and up to the time of impact with the crash attenuator, the Tesla's speed increased from 62 to 70,8 mph, with no precrash braking or evasive steering movement detected." Devin Coldewey, "Uber in fatal crash detected pedestrian but had emergency braking disabled," *TechCrunch*, May 24, 2018: "Emergency braking maneuvers are not enabled while the vehicle is under computer control, to reduce the potential for erratic vehicle behavior."

(6) The Society of Automotive Engineers (SAE) defines six levels of automation, where Level 0 is none at all and Level 5 is full automation: "The full-time performance by an automatic driving system of all aspects of the dynamic driving task under all roadway and environmental conditions that can be managed by a human driver."

(7) Forecast of economic effects of automation on transportation costs: Adele Peters, “It could be 10 times cheaper to take electric robo-taxis than to own a car by 2030,” *Fast Company*, May 30, 2017.

(8) The impact of accidents on the prospects for regulatory action on autonomous vehicles: Richard Waters, “Self-driving car death poses dilemma for regulators,” *Financial Times*, March 20, 2018.

(9) The impact of accidents on public perception of autonomous vehicles: Cox Automotive, “Autonomous vehicle awareness rising, acceptance declining, according to Cox Automotive mobility study,” August 16, 2018.

(10) The original chatbot: Joseph Weizenbaum, “ELIZA — a computer program for the study of natural language communication between man and machine,” *Communications of the ACM* 9 (1966): 36–45.

(11) See physiome.org for current activities in physiological modeling. Work in the 1960s assembled models with thousands of differential equations: Arthur Guyton, Thomas Coleman, and Harris Granger, “Circulation: Overall regulation,” *Annual Review of Physiology* 34 (1972): 13–44.

(12) Some of the earliest work on tutoring systems was done by Pat Suppes and colleagues at Stanford: Patrick Suppes and Mona Morningstar, “Computer-assisted instruction,” *Science* 166 (1969): 343–50.

(13) Michael Yudelson, Kenneth Koedinger, and Geoffrey Gordon, “Individualized Bayesian knowledge tracing models,” in *Artificial Intelligence in Education: 16th International Conference*, ed. H. Chad Lane et al. (Springer, 2013).

(14) For an example of machine learning on encrypted data, see, for example, Reza Shokri and Vitaly Shmatikov, “Privacy-preserving deep learning,” in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security* (ACM, 2015).

(15) A retrospective on the first smart home, based on a lecture by its inventor, James Sutherland: James E. Tomayko, “Electronic Computer

for Home Operation (ECHO): The first home computer," *IEEE Annals of the History of Computing* 16 (1994): 59–61.

(16) Summary of a smart-home project based on machine learning and automated decisions: Diane Cook et al., "MavHome: An agent-based smart home," in *Proceedings of the 1st IEEE International Conference on Pervasive Computing and Communications* (IEEE, 2003).

(17) For the beginnings of an analysis of user experiences in smart homes, see Scott Davidoff et al., "Principles of smart home control," in *Ubicomp 2006: Ubiquitous Computing*, ed. Paul Dourish and Adrian Friday (Springer, 2006).

(18) Commercial announcement of AI-based smart homes: "The Wolff Company unveils revolutionary smart home technology at new Annadel Apartments in Santa Rosa, California," *Business Insider*, March 12, 2018.

(19) Article on robot chefs as commercial products: Eustacia Huen, "The world's first home robotic chef can cook over 100 meals," *Forbes*, October 31, 2016.

(20) Report from my Berkeley colleagues on deep RL for robotic motor control: Sergey Levine et al., "End-to-end training of deep visuomotor policies," *Journal of Machine Learning Research* 17 (2016): 1–40.

(21) On the possibilities for automating the work of hundreds of thousands of warehouse workers: Tom Simonite, "Grasping robots compete to rule Amazon's warehouses," *Wired*, July 26, 2017.

(22) I'm assuming a generous one laptop-CPU minute per page, or about 10^{11} operations. A third-generation tensor processing unit from Google runs at about 10^{17} operations per second, meaning that it can read a million pages per second, or about five hours for eighty million two-hundred-page books.

(23) A 2003 study on the global volume of information production by all channels: Peter Lyman and Hal Varian, “How much information?” sims.berkeley.edu/research/projects/how-much-info-2003.

(24) For details on the use of speech recognition by intelligence agencies, see Dan Froomkin, “How the NSA converts spoken words into searchable text,” *The Intercept*, May 5, 2015.

(25) Analysis of visual imagery from satellites is an enormous task: Mike Kim, “Mapping poverty from space with the World Bank,” Medium.com, January 4, 2017. Kim estimates eight million people working 24/7, which converts to more than thirty million people working forty hours per week. I suspect this is an overestimate in practice, because the vast majority of the images would exhibit negligible change over the course of one day. On the other hand, the US intelligence community employs tens of thousands of people sitting in vast rooms staring at satellite images just to keep track of what’s happening in small regions of interest; so one million people is probably about right for the whole world.

(26) There is substantial progress towards a global observatory based on real-time satellite image data: David Jensen and Jillian Campbell, “Digital earth: Building, financing and governing a digital ecosystem for planetary data,” white paper for the UN Science-Policy-Business Forum on the Environment, 2018.

(27) Luke Muehlhauser has written extensively on AI predictions, and I am indebted to him for tracking down original sources for the quotations that follow. See Luke Muehlhauser, “What should we learn from past AI forecasts?” Open Philanthropy Project report, 2016.

(28) A forecast of the arrival of human-level AI within twenty years: Herbert Simon, *The New Science of Management Decision* (Harper & Row, 1960).

(29) A forecast of the arrival of human-level AI within a generation: Marvin Minsky, *Computation: Finite and Infinite Machines* (Prentice Hall, 1967).

(30) John McCarthy's forecast of the arrival of human-level AI within "five to 500 years": Ian Shenker, "Brainy robots in our future, experts think," *Detroit Free Press*, September 30, 1977.

(31) For a summary of surveys of AI researchers on their estimates for the arrival of humanlevel AI, see aiimpacts.org. An extended discussion of survey results on human-level AI is given by Katja Grace et al., "When will AI exceed human performance? Evidence from AI experts," arXiv:1705.08807v3 (2018).

(32) For a chart mapping raw computer power against brain power, see Ray Kurzweil, "The law of accelerating returns," Kurzweilai.net, March 7, 2001.

(33) The Allen Institute's Project Aristo: allenai.org/aristo.

(34) For an analysis of the knowledge required to perform well on fourth-grade tests of comprehension and common sense, see Peter Clark et al., "Automatic construction of inference-supporting knowledge bases," in *Proceedings of the Workshop on Automated Knowledge Base Construction (*2014), akbc.ws/2014.

(35) The NELL project on machine reading is described by Tom Mitchell et al., "Neverending learning," *Communications of the ACM* 61 (2018): 103–15.

(36) The idea of bootstrapping inferences from text is due to Sergey Brin, "Extracting patterns and relations from the World Wide Web," in *The World Wide Web and Databases*, ed. Paolo Atzeni, Alberto Mendelzon, and Giansalvatore Mecca (Springer, 1998).

(37) For a visualization of the black-hole collision detected by LIGO, see LIGO Lab Caltech, “Warped space and time around colliding black holes,” February 11, 2016, youtube.com/watch?v=1agm33iEAuo.

(38) The first publication describing observation of gravitational waves: Abbott et al., “Observation of gravitational waves from a binary black hole *Physical Review Letters* 116 (2016): 061102.

(39) On babies as scientists: Alison Gopnik, Andrew Meltzoff, Patricia Kuhl, *The Scientist in the Crib: Minds, Brains, and How Children Learn* (William Morrow, 1999).

(40) A summary of several projects on automated scientific analysis of experimental data to discover laws: Patrick Langley et al., *Scientific Discovery: Computational Explorations of the Creative Processes* (MIT Press, 1987).

(41) Some early work on machine learning guided by prior knowledge: Stuart Russell, *The Use of Knowledge in Analogy and Induction* (Pitman, 1989).

(42) Goodman’s philosophical analysis of induction remains a source of inspiration: Nelson Goodman, *Fact, Fiction, and Forecast* (University of London Press, 1954).

(43) A veteran AI researcher complains about mysticism in the philosophy of science: Herbert Simon, “Explaining the ineffable: AI on the topics of intuition, insight and inspiration,” in *Proceedings of the 14th International Conference on Artificial Intelligence*, ed. Chris Mellish (Morgan Kaufmann, 1995).

(44) A survey of inductive logic programming by two originators of the field: Stephen Muggleton and Luc de Raedt, “Inductive logic programming: Theory and methods,” *Journal of Logic Programming* 19-20 (1994): 629–79.

(45) For an early mention of the importance of encapsulating complex operations as new primitive actions, see Alfred North Whitehead, *An Introduction to Mathematics* (Henry Holt, 1911).

(46) Work demonstrating that a simulated robot can learn entirely by itself to stand up: John Schulman et al., "High-dimensional continuous control using generalized advantage estimation," arXiv:1506.02438 (2015). A video demonstration is available at youtube.com/watch?v=SHLuf2ZBQSw.

(47) A description of a reinforcement learning system that learns to play a capture-the-flag video game: Max Jaderberg et al., "Human-level performance in first-person multiplayer games with population-based deep reinforcement learning," arXiv:1807.01281 (2018).

(48) A view of AI progress over the next few years: Peter Stone et al., "Artificial intelligence and life in 2030," *One Hundred Year Study on Artificial Intelligence*, report of the 2015 Study Panel, 2016.

(49) The media-fueled argument between Elon Musk and Mark Zuckerberg: Peter Holley, "Billionaire burn: Musk says Zuckerberg's understanding of AI threat 'is limited,'" *The Washington Post*, July 25, 2017.

(50) On the value of search engines to individual users: Erik Brynjolfsson, Felix Eggers, and Avinash Gannamaneni, "Using massive online choice experiments to measure changes in well-being," working paper no. 24514, National Bureau of Economic Research, 2018.

(51) Penicillin was discovered several times and its curative powers were described in medical publications, but no one seems to have noticed. See en.wikipedia.org/wiki/History_of_penicillin.

(52) For a discussion of some of the more esoteric risks from omniscient, clairvoyant AI systems, see David Auerbach, "The most terrifying thought experiment of all time," *Slate*, July 17, 2014.

(53) An analysis of some potential pitfalls in thinking about advanced AI: Kevin Kelly, "The myth of a superhuman AI," *Wired*, April 25, 2017.

(54) Machines may share *some* aspects of cognitive structure with humans, particularly those aspects dealing with perception and manipulation of the physical world and the conceptual structures involved in natural language understanding. Their deliberative processes are likely to be quite different because of the enormous disparities in hardware.

(55) According to 2016 survey data, the eighty-eighth percentile corresponds to $100,000 per year: American Community Survey, US Census Bureau, www.census.gov/programs-surveys/acs. For the same year, global per capita GDP was $10,133: National Accounts Main Aggregates Database, UN Statistics Division, unstats.un.org/unsd/snaama.

(56) If the GDP growth phases in over ten years or twenty years, it's worth $9,400 trillion or $6,800 trillion, respectively — still nothing to sneeze at. On an interesting historical note, I. J. Good, who popularized the notion of an intelligence explosion (page 142), estimated the value of human-level AI to be at least "one megaKeynes," referring to the fabled economist John Maynard Keynes. The value of Keynes's contributions was estimated in 1963 as £100 billion, so a megaKeynes comes out to around $2,200,000 trillion in 2016 dollars. Good pinned the value of AI primarily on its potential to ensure that the human race survives indefinitely. Later, he came to wonder whether he should have added a minus sign.

(57) The EU announced plans for $24 billion in research and development spending for the period 2019–20. See European Commission, "Artificial intelligence: Commission outlines a European approach to boost investment and set ethical guidelines," press release, April 25, 2018. China's long-term investment plan for AI, announced in 2017, envisages a core AI industry generating $150 billion annually by 2030. See, for example, Paul

Mozur, “Beijing wants A.I. to be made in China by 2030,” *The New York Times*, July 20, 2017.

(58) See, for example, Rio Tinto’s Mine of the Future program at riotinto.com/australia/pilbara/mine-of-the-future-9603.aspx.

(59) A retrospective analysis of economic growth: Jan Luiten van Zanden et al., eds., *How Was Life? Global Well-Being since 1820* (OECD Publishing, 2014).

(60) The desire for relative advantage over others, rather than an absolute quality of life, is a *positional good*; see Chapter 9.

الفصل الرابع: إساءة استخدام الذكاء الاصطناعي

(1) Wikipedia’s article on the Stasi has several useful references on its workforce and its overall impact on East German life.

(2) For details on Stasi files, see Cullen Murphy, *God’s Jury: The Inquisition and the Making of the Modern World* (Houghton Mifflin Harcourt, 2012).

(3) For a thorough analysis of AI surveillance systems, see Jay Stanley, *The Dawn of Robot Surveillance* (American Civil Liberties Union, 2019).

(4) Recent books on surveillance and control include Shoshana Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (PublicAffairs, 2019) and Roger McNamee, *Zucked: Waking Up to the Facebook Catastrophe* (Penguin Press, 2019).

(5) News article on a blackmail bot: Avivah Litan, “Meet Delilah — the first insider threat Trojan,” Gartner Blog Network, July 14, 2016.

(6) For a low-tech version of human susceptibility to misinformation, in which an unsuspecting individual becomes convinced that the world is being destroyed by meteor strikes, see *Derren Brown: Apocalypse*, “Part One,” directed by Simon Dinsell, 2012, youtube.com/watch?v=o_CUrM JOxqs.

(7) An economic analysis of reputation systems and their corruption is given by Steven Tadelis, “Reputation and feedback systems in online platform markets,” *Annual Review of Economics* 8 (2016): 321–40.

(8) Goodhart’s law: “Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes.” For example, there may once have been a correlation between faculty quality and faculty salary, so the *US News & World Report* college rankings measure faculty quality by faculty salaries. This has contributed to a salary arms race that benefits faculty members but not the students who pay for those salaries. The arms race changes faculty salaries in a way that does not depend on faculty quality, so the correlation tends to disappear.

(9) An article describing German efforts to police public discourse: Bernhard Rohleder, “Germany set out to delete hate speech online. Instead, it made things worse,” *World-Post*, February 20, 2018.

(10) On the “infopocalypse”: Aviv Ovadya, “What’s worse than fake news? The distortion of reality itself,” *WorldPost*, February 22, 2018.

(11) On the corruption of online hotel reviews: Dina Mayzlin, Yaniv Dover, and Judith Chevalier, “Promotional reviews: An empirical investigation review manipulation,” *American Economic Review* 104 (2014): 2421–55.

(12) Statement of Germany at the Meeting of the Group of Governmental Experts, Convention on Certain Conventional Weapons, Geneva, April 10, 2018.

(13) The *Slaughterbots* movie, funded by the Future of Life Institute, appeared in November 2017 and is available at youtube.com/watch?v=9CO6M2HsoIA.

(14) For a report on one of the bigger *faux pas* in military public relations, see Dan Lamothe, “Pentagon agency wants drones to hunt in packs, like wolves,” *The Washington Post*, January 23, 2015.

(15) Announcement of a large-scale drone swarm experiment: US Department of Defense, “Department of Defense announces successful micro-drone demonstration,” news release no. NR-008-17, January 9, 2017.

(16) Examples of research centers studying the impact of technology on employment are the Work and Intelligent Tools and Systems group at Berkeley, the Future of Work and Workers project at the Center for Advanced Study in the Behavioral Sciences at Stanford, and the Future of Work Initiative at Carnegie Mellon University.

(17) A pessimistic take on future technological unemployment: Martin Ford, *Rise of the Robots: Technology and the Threat of a Jobless Future* (Basic Books, 2015).

(18) Calum Chace, *The Economic Singularity: Artificial Intelligence and the Death of Capitalism* (Three Cs, 2016).

(19) For an excellent collection of essays, see Ajay Agrawal, Joshua Gans, and Avi Goldfarb, eds., *The Economics of Artificial Intelligence: An Agenda* (National Bureau of Economic Research, 2019).

(20) The mathematical analysis behind this “inverted-U” employment curve is given by James Bessen, “Artificial intelligence and jobs: The role of demand” in *The Economics of Artificial Intelligence*, ed. Agrawal, Gans, and Goldfarb.

(21) For a discussion of economic dislocation arising from automation, see Eduardo Porter, “Tech is splitting the US work force in two,” *The New York Times*, February 4, 2019. The article cites the following report for this conclusion: David Autor and Anna Salomons, “Is automation labor-displacing? Productivity growth, employment, and the labor share,” *Brookings Papers on Economic Activity* (2018).

(22) For data on the growth of banking in the twentieth century, see Thomas Philippon, "The evolution of the US financial industry from 1860 to 2007: Theory and evidence," working paper, 2008.

(23) The bible for jobs data and the growth and decline of occupations: US Bureau of Labor Statistics, *Occupational Outlook Handbook: 2018–2019 Edition* (Bernan Press, 2018).

(24) A report on trucking automation: Lora Kolodny, "Amazon is hauling cargo in selfdriving trucks developed by Embark," CNBC, January 30, 2019.

(25) The progress of automation in legal analytics, describing the results of a contest: Jason Tashea, "AI software is more accurate, faster than attorneys when assessing NDAs," *ABA Journal*, February 26, 2018.

(26) A commentary by a distinguished economist, with a title explicitly evoking Keynes's 1930 article: Lawrence Summers, "Economic possibilities for our children," *NBER Reporter* (2013).

(27) The analogy between data science employment and a small lifeboat for a giant cruise ship comes from a discussion with Yong Ying-I, head of Singapore's Public Service Division. She conceded that it was correct on the global scale, but noted that "Singapore is small enough to fit in the lifeboat."

(28) Support for UBI from a conservative viewpoint: Sam Bowman, "The ideal welfare system is a basic income," Adam Smith Institute, November 25, 2013.

(29) Support for UBI from a progressive viewpoint: Jonathan Bartley, "The Greens endorse a universal basic income. Others need to follow," *The Guardian*, June 2, 2017.

(30) Chace, in *The Economic Singularity*, calls the "paradise" version of UBI the *Star Trek economy,* noting that in the more recent series of *Star Trek* episodes, money has been abolished because technology has created

essentially unlimited material goods and energy. He also points to the massive changes in economic and social organization that will be needed to make such a system successful.

(31) The economist Richard Baldwin also predicts a future of personal services in his book *The Globotics Upheaval: Globalization, Robotics, and the Future of Work* (Oxford University Press, 2019).

(32) The book that is viewed as having exposed the failure of "whole-word" literacy education and launched decades of struggle between the two main schools of thought on reading: Rudolf Flesch, *Why Johnny Can't Read: And What You Can Do about It* (Harper & Bros., 1955).

(33) On educational methods that enable the recipient to adapt to the rapid rate of technological and economic change in the next few decades: Joseph Aoun, *Robot-Proof: Higher Education in the Age of Artificial Intelligence* (MIT Press, 2017).

(34) A radio lecture in which Turing predicted that humans would be overtaken by machines: Alan Turing, "Can digital machines think?," May 15, 1951, radio broadcast, BBC Third Programme. Typescript available at turingarchive.org.

(35) News article describing the "naturalization" of Sophia as a citizen of Saudi Arabia: Dave Gershgorn, "Inside the mechanical brain of the world's first robot citizen," *Quartz*, November 12, 2017.

(36) On Yann LeCun's view of Sophia: Shona Ghosh, "Facebook's AI boss described Sophia the robot as 'complete b——t' and 'Wizard-of-Oz AI,'" *Business Insider*, January 6, 2018.

(37) An EU proposal on legal rights for robots: Committee on Legal Affairs of the European Parliament, "Report with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL))," 2017.

(38) The GDPR provision on a "right to an explanation" is not, in fact, new: it is very similar to Article 15(1) of the 1995 Data Protection Directive, which it supersedes.

(39) Here are three recent papers providing insightful mathematical analyses of fairness: Moritz Hardt, Eric Price, and Nati Srebro, "Equality of opportunity in supervised learning," in *Advances in Neural Information Processing Systems 29*, ed. Daniel Lee et al. (2016); Matt Kusner et al., "Counterfactual fairness," in *Advances in Neural Information Processing Systems 30*, ed. Isabelle Guyon et al. (2017); Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan, "Inherent trade-offs in the fair determination of risk scores," in *8th Innovations in Theoretical Computer Science Conference*, ed. Christos Papadimitriou (Dagstuhl Publishing, 2017).

(40) News article describing the consequences of software failure for air traffic control: Simon Calder, "Thousands stranded by flight cancellations after systems failure at Europe's air-traffic coordinator," *The Independent*, April 3, 2018.

الفصل الخامس: الذكاء الاصطناعي الفائق الذكاء

(1) Lovelace wrote, "The Analytical Engine has no pretensions whatever to originate anything. It can do whatever we know how to order it to perform. It can follow analysis; but it has no power of anticipating any analytical relations or truths." This was one of the arguments against AI that was refuted by Alan Turing, "Computing machinery and intelligence," *Mind* 59 (1950): 433–60.

(2) The earliest known article on existential risk from AI was by Richard Thornton, "The age of machinery," *Primitive Expounder* IV (1847): 281.

(3) "The Book of the Machines" was based on an earlier article by Samuel Butler, "Darwin among the machines," *The Press* (Christchurch, New Zealand), June 13, 1863.

(4) Another lecture in which Turing predicted the subjugation of humankind: Alan Turing, "Intelligent machinery, a heretical theory" (lecture given to the 51 Society, Manchester, 1951). Typescript available at turingarchive.org.

(5) Wiener's prescient discussion of technological control over humanity and a plea to retain human autonomy: Norbert Wiener, *The Human Use of Human Beings* (Riverside Press, 1950).

(6) The front-cover blurb from Wiener's 1950 book is remarkably similar to the motto of the Future of Life Institute, an organization dedicated to studying the existential risks that humanity faces: "Technology is giving life the potential to flourish like never before … or to self-destruct."

(7) An updating of Wiener's views arising from his increased appreciation of the possibility of intelligent machines: Norbert Wiener, *God and Golem, Inc.: A Comment on Certain Points Where Cybernetics Impinges on Religion* (MIT Press, 1964).

(8) Asimov's Three Laws of Robotics first appeared in Isaac Asimov, "Runaround," *Astounding Science Fiction*, March 1942. The laws are as follows:

(1) A robot may not injure a human being or, through inaction, allow a human being to come to harm.

(2) A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.

(3) A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.

It is important to understand that Asimov proposed these laws as a way to generate interesting story plots, not as a serious guide for future roboticists. Several of his stories, including “Runaround,” illustrate the problematic consequences of taking the laws literally. From the standpoint of modern AI, the laws fail to acknowledge any element of probability and risk: the legality of robot actions that expose a human to some probability of harm — however infinitesimal — is therefore unclear.

(9) The notion of instrumental goals is due to Stephen Omohundro, “The nature of selfimproving artificial intelligence” (unpublished manuscript, 2008). See also Stephen Omohundro, “The basic AI drives,” in *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, ed. Pei Wang, Ben Goertzel, and Stan Franklin (IOS Press, 2008).

(10) The objective of Johnny Depp’s character, Will Caster, seems to be to solve the problem of physical reincarnation so that he can be reunited with his wife, Evelyn. This just goes to show that the nature of the overarching objective doesn’t matter — the instrumental goals are all the same.

(11) The original source for the idea of an intelligence explosion: I. J. Good, “Speculations concerning the first ultraintelligent machine,” in *Advances in Computers*, vol. 6, ed. Franz Alt and Morris Rubinoff (Academic Press, 1965).

(12) An example of the impact of the intelligence explosion idea: Luke Muehlhauser, in *Facing the Intelligence Explosion* (intelligenceexplosion.com), writes, “Good’s paragraph ran over me like a train.”

(13) Diminishing returns can be illustrated as follows: suppose that a 16 percent improvement in intelligence creates a machine capable of making an 8 percent improvement, which in turn creates a 4 percent improvement, and so on. This process reaches a limit at about 36 percent above the original level. For more discussion on these issues,

see Eliezer Yudkowsky, “Intelligence explosion microeconomics,” technical report 2013-1, Machine Intelligence Research Institute, 2013.

(14) For a view of AI in which humans become irrelevant, see Hans Moravec, *Mind Children: The Future of Robot and Human Intelligence* (Harvard University Press, 1988). See also Hans Moravec, *Robot: Mere Machine to Transcendent Mind* (Oxford University Press, 2000).

الفصل السادس: الجدل غير الواسع الدائر حول الذكاء الاصطناعي

(1) A serious publication provides a serious review of Bostrom’s *Superintelligence: Paths, Dangers, Strategies*: “Clever cogs,” *Economist*, August 9, 2014.

(2) A discussion of myths and misunderstandings concerning the risks of AI: Scott Alexander, “AI researchers on AI risk,” *Slate Star Codex* (blog), May 22, 2015.

(3) The classic work on multiple dimensions of intelligence: Howard Gardner, *Frames of Mind: The Theory of Multiple Intelligences* (Basic Books, 1983).

(4) On the implications of multiple dimensions of intelligence for the possibility of superhuman AI: Kevin Kelly, “The myth of a superhuman AI,” *Wired*, April 25, 2017.

(5) Evidence that chimpanzees have better short-term memory than humans: Sana Inoue and Tetsuro Matsuzawa, “Working memory of numerals in chimpanzees,” *Current Biology* 17 (2007), R1004–5.

(6) An important early work questioning the prospects for rule-based AI systems: Hubert Dreyfus, *What Computers Can’t Do* (MIT Press, 1972).

(7) The first in a series of books seeking physical explanations for consciousness and raising doubts about the ability of AI systems to achieve real intelligence: Roger Penrose, *The Emperor’s New Mind: Concerning Computers, Minds, and the Laws of Physics* (Oxford University Press, 1989).

(8) A revival of the critique of AI based on the incompleteness theorem: Luciano Floridi, “Should we be afraid of AI?” *Aeon*, May 9, 2016.

(9) A revival of the critique of AI based on the Chinese room argument: John Searle, “What your computer can’t know,” *The New York Review of Books*, October 9, 2014.

(10) A report from distinguished AI researchers claiming that superhuman AI is probably impossible: Peter Stone et al., “Artificial intelligence and life in 2030,” One Hundred Year Study on Artificial Intelligence, report of the 2015 Study Panel, 2016.

(11) News article based on Andrew Ng’s dismissal of risks from AI: Chris Williams, “AI guru Ng: Fearing a rise of killer robots is like worrying about overpopulation on Mars,” *Register*, March 19, 2015.

(12) An example of the “experts know best” argument: Oren Etzioni, “It’s time to intelligently discuss artificial intelligence,” *Backchannel*, December 9, 2014.

(13) News article claiming that real AI researchers dismiss talk of risks: Erik Sofge, “Bill Gates fears AI, but AI researchers know better,” *Popular Science*, January 30, 2015.

(14) Another claim that real AI researchers dismiss AI risks: David Kenny, “IBM’s open letter to Congress on artificial intelligence,” June 27, 2017, ibm.com/blogs/policy/kenny-artificial-intelligence-letter.

(15) Report from the workshop that proposed voluntary restrictions on genetic engineering: Paul Berg et al., “Summary statement of the Asilomar Conference on Recombinant DNA Molecules,” *Proceedings of the National Academy of Sciences* 72 (1975): 1981–84.

(16) Policy statement arising from the invention of CRISPR-Cas9 for gene editing: Organizing Committee for the International Summit on Human Gene Editing, “On human gene editing: International Summit statement,” December 3, 2015.

(17) The latest policy statement from leading biologists: Eric Lander et al., “Adopt a moratorium on heritable genome editing,” *Nature* 567 (2019): 165–68.

(18) Etzioni’s comment that one cannot mention risks if one does not also mention benefits appears alongside his analysis of survey data from AI researchers: Oren Etzioni, “No, the experts don’t think superintelligent AI is a threat to humanity,” *MIT Technology Review*, September 20, 2016. In his analysis he argues that anyone who expects superhuman AI to take more than twenty-five years — which includes this author as well as Nick Bostrom — is not concerned about the risks of AI.

(19) A news article with quotations from the Musk-Zuckerberg “debate”: Alanna Petroff, “Elon Musk says Mark Zuckerberg’s understanding of AI is ‘limited,’” *CNN Money*, July 25, 2017.

(20) In 2015 the Information Technology and Innovation Foundation organized a debate titled “Are super intelligent computers really a threat to humanity?” Robert Atkinson, director of the foundation, suggests that mentioning risks is likely to result in reduced funding for AI. Video available at itif.org/events/2015/06/30/are-super-intelligent-computers-really-threat-humanity; the relevant discussion begins at 41:30.

(21) A claim that our culture of safety will solve the AI control problem without ever mentioning it: Steven Pinker, “Tech prophecy and the underappreciated causal power of ideas,” in *Possible Minds: Twenty-Five Ways of Looking at AI*, ed. John Brockman (Penguin Press, 2019).

(22) For an interesting analysis of Oracle AI, see Stuart Armstrong, Anders Sandberg, and Nick Bostrom, “Thinking inside the box: Controlling and using an Oracle AI,” *Minds and Machines* 22 (2012): 299–324.

(23) Views on why AI is not going to take away jobs: Kenny, “IBM’s open letter.”

(24) An example of Kurzweil's positive views of merging human brains with AI: Ray Kurzweil, interview by Bob Pisani, June 5, 2015, Exponential Finance Summit, New York, NY.

(25) Article quoting Elon Musk on neural lace: Tim Urban, "Neuralink and the brain's magical future," Wait But Why, April 20, 2017.

(26) For the most recent developments in Berkeley's neural dust project, see David Piech et al., "StimDust: A 1,7 mm^3, implantable wireless precision neural stimulator with ultrasonic power and communication," arXiv: 1807.07590 (2018).

(27) Susan Schneider, in *Artificial You: AI and the Future of Your Mind* (Princeton University Press, 2019), points out the risks of ignorance in proposed technologies such as uploading and neural prostheses: that, absent any real understanding of whether electronic devices can be conscious and given the continuing philosophical confusion over persistent personal identity, we may inadvertently end our own conscious existences or inflict suffering on conscious machines without realizing that they are conscious.

(28) An interview with Yann LeCun on AI risks: Guia Marie Del Prado, "Here's what Facebook's artificial intelligence expert thinks about the future," *Business Insider*, September 23, 2015.

(29) A diagnosis of AI control problems arising from an excess of testosterone: Steven Pinker, "Thinking does not imply subjugating," in *What to Think About Machines That Think*, ed. John Brockman (Harper Perennial, 2015).

(30) A seminal work on many philosophical topics, including the question of whether moral obligations may be perceived in the natural world: David Hume, *A Treatise of Human Nature* (John Noon, 1738).

(31) An argument that a sufficiently intelligent machine cannot help but pursue human objectives: Rodney Brooks, "The seven deadly sins of AI predictions," *MIT Technology Review*, October 6, 2017.

(32) Pinker, "Thinking does not imply subjugating."

(33) For an optimistic view arguing that AI safety problems will necessarily be resolved in our favor: Steven Pinker, "Tech prophecy."

(34) On the unsuspected alignment between "skeptics" and "believers" in AI risk: Alexander, "AI researchers on AI risk."

الفصل السابع: الذكاء الاصطناعي: توجُّه مُختلف

(1) For a guide to detailed brain modeling, now slightly outdated, see Anders Sandberg and Nick Bostrom, "*Whole brain emulation: A roadmap*," technical report 2008–3, Future of Humanity Institute, Oxford University, 2008.

(2) For an introduction to genetic programming from a leading exponent, see John Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection* (MIT Press, 1992).

(3) The parallel to Asimov's Three Laws of Robotics is entirely coincidental.

(4) The same point is made by Eliezer Yudkowsky, "Coherent extrapolated volition," technical report, Singularity Institute, 2004. Yudkowsky argues that directly building in "Four Great Moral Principles That Are All We Need to Program into AIs" is a sure road to ruin for humanity. His notion of the "coherent extrapolated volition of humankind" has the same general flavor as the first principle; the idea is that a superintelligent AI system could work out what humans, collectively, really want.

(5) You can certainly have preferences over whether a machine is helping you achieve your preferences or you are achieving them through your own efforts. For example, suppose you prefer outcome A to outcome B, all other things being equal. You are unable to achieve outcome A unaided, and yet you still prefer B to getting A with the machine's help. In that case

the machine should decide not to help you — unless perhaps it can do so in a way that is completely undetectable by you. You may, of course, have preferences about undetectable help as well as detectable help.

(6) The phrase "the greatest good of the greatest number" originates in the work of Francis Hutcheson, *An Inquiry into the Original of Our Ideas of Beauty and Virtue, In Two Treatises* (D. Midwinter et al., 1725). Some have ascribed the formulation to an earlier comment by Wilhelm Leibniz; see Joachim Hruschka, "The greatest happiness principle and other early German anticipations of utilitarian theory," *Utilitas* 3 (1991): 165–77.

(7) One might propose that the machine should include terms for animals as well as humans in its own objective function. If these terms have weights that correspond to how much people care about animals, then the end result will be the same as if the machine cares about animals only through caring about humans who care about animals. Giving each living animal equal weight in the machine's objective function would certainly be catastrophic — for example, we are outnumbered fifty thousand to one by Antarctic krill and a billion trillion to one by bacteria.

(8) The moral philosopher Toby Ord made the same point to me in his comments on an early draft of this book: "Interestingly, the same is true in the study of moral philosophy. Uncertainty about moral value of outcomes was almost completely neglected in moral philosophy until very recently. Despite the fact that it is our uncertainty of moral matters that leads people to ask others for moral advice and, indeed, to do research on moral philosophy at all!"

(9) One excuse for not paying attention to uncertainty about preferences is that it is formally equivalent to ordinary uncertainty, in the following sense: being uncertain about what I like is the same as being certain that I like likable things while being uncertain about what things are likable.

This is just a trick that appears to move the uncertainty into the world, by making "likability by me" a property of objects rather than a property of me. In game theory, this trick has been thoroughly institutionalized since the 1960s, following a series of papers by my late colleague and Nobel laureate John Harsanyi: "Games with incomplete information played by 'Bayesian' players, Parts I–III," *Management Science* 14 (1967, 1968): 159–82, 320–34, 486–502. In decision theory, the standard reference is the following: Richard Cyert and Morris de Groot, "Adaptive utility," in *Expected Utility Hypotheses and the Allais Paradox*, ed. Maurice Allais and Ole Hagen (D. Reidel, 1979).

(10) AI researchers working in the area of preference elicitation are an obvious exception. See, for example, Craig Boutilier, "On the foundations of *expected* expected utility," in *Proceedings of the 18th International Joint Conference on Artificial Intelligen*ce (Morgan Kaufmann, 2003). Also Alan Fern et al., "A decision-theoretic model of assistance," *Journal of Artificial Intelligence Research* 50 (2014): 71–104.

(11) A critique of beneficial AI based on a misinterpretation of a journalist's brief interview with the author in a magazine article: Adam Elkus, "How to be good: Why you can't teach human values to artificial intelligence," *Slate*, April 20, 2016.

(12) The origin of trolley problems: Frank Sharp, "A study of the influence of custom on the moral judgment," *Bulletin of the University of Wisconsin* 236 (1908).

(13) The "anti-natalist" movement believes it is morally wrong for humans to reproduce because to live is to suffer and because humans' impact on the Earth is profoundly negative. If you consider the existence of humanity to be a moral dilemma, then I suppose I do want machines to resolve this moral dilemma the right way.

(14) Statement on China's AI policy by Fu Ying, vice chair of the Foreign Affairs Committee of the National People's Congress. In a letter to the 2018 World AI Conference in Shanghai, Chinese president Xi Jinping wrote, "Deepened international cooperation is required to cope with new issues in fields including law, security, employment, ethics and governance." I am indebted to Brian Tse for bringing these statements to my attention.

(15) A very interesting paper on the non-naturalistic non-fallacy, showing how preferences can be inferred from the state of the world as arranged by humans: Rohin Shah et al., "The implicit preference information in an initial state," in *Proceedings of the 7th International Conference on Learning Representations* (2019), iclr.cc/Conferences/2019/Schedule.

(16) Retrospective on Asilomar: Paul Berg, "Asilomar 1975: DNA modification secured," *Nature* 455 (2008): 290–91.

(17) News article reporting Putin's speech on AI: "Putin: Leader in artificial intelligence will rule world," Associated Press, September 4, 2017.

الفصل الثامن: الذكاء الاصطناعي النافع على نحو مثبت

(1) Fermat's Last Theorem asserts that the equation $a^n = b^n + C^n$ has no solutions with *a*, *b*, and *c* being whole numbers and *n* being a whole number larger than 2. In the margin of his copy of Diophantus's *Arithmetica*, Fermat wrote, "I have a truly marvellous proof of this proposition which this margin is too narrow to contain." True or not, this guaranteed that mathematicians pursued a proof with vigor in the subsequent centuries. We can easily check particular cases — for example, is 7^3 equal to $6^3 + 5^3$? (Almost, because 7^3 is 343 and $6^3 + 5^3$ is 341, but "almost" doesn't count.) There are, of course, infinitely many cases to check, and that's why we need mathematicians and not just computer programmers.

(2) A paper from the Machine Intelligence Research Institute poses many related issues: Scott Garrabrant and Abram Demski, “Embedded agency,” AI Alignment Forum, November 15, 2018.

(3) The classic work on multiattribute utility theory: Ralph Keeney and Howard Raiffa, *Decisions with Multiple Objectives: Preferences and Value Tradeoffs* (Wiley, 1976).

(4) Paper introducing the idea of inverse RL: Stuart Russell, “Learning agents for uncertain environments,” in *Proceedings of the 11th Annual Conference on Computational Learning Theory* (ACM, 1998).

(5) The original paper on structural estimation of Markov decision processes: Thomas Sargent, “Estimation of dynamic labor demand schedules under rational expectations,” *Journal of Political Economy* 86 (1978): 1009–44.

(6) The first algorithms for IRL: Andrew Ng and Stuart Russell, “Algorithms for inverse reinforcement learning,” in *Proceedings of the 17th International Conference on Machine Learning*, ed. Pat Langley (Morgan Kaufmann, 2000).

(7) Better algorithms for inverse RL: Pieter Abbeel and Andrew Ng, “Apprenticeship learning via inverse reinforcement learning,” in *Proceedings of the 21st International Conference on Machine Learning*, ed. Russ Greiner and Dale Schuurmans (ACM Press, 2004).

(8) Understanding inverse RL as Bayesian updating: Deepak Ramachandran and Eyal Amir, “Bayesian inverse reinforcement learning,” in *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, ed. Manuela Veloso (AAAI Press, 2007)

(9) How to teach helicopters to fly and do aerobatic maneuvers: Adam Coates, Pieter Abbeel, and Andrew Ng, “Apprenticeship learning for helicopter control,” *Communications of the ACM* 52 (2009): 97–105.

(10) The original name proposed for an assistance game was a *co-operative inverse reinforcement learning* game, or CIRL game. See Dylan Hadfield-Menell et al., "Cooperative inverse reinforcement learning," in *Advances in Neural Information Processing Systems 29*, ed. Daniel Lee et al. (2016).

(11) These numbers are chosen just to make the game interesting.

(12) The equilibrium solution to the game can be found by a process called *iterated best response*: pick any strategy for Harriet; pick the best strategy for Robbie, given Harriet's strategy; pick the best strategy for Harriet, given Robbie's strategy; and so on. If this process reaches a fixed point, where neither strategy changes, then we have found a solution. The process unfolds as follows:

(1) Start with the greedy strategy for Harriet: make 2 paperclips if she prefers paperclips; make 1 of each if she is indifferent; make 2 staples if she prefers staples.

(2) There are three possibilities Robbie has to consider, given this strategy for Harriet:

(a) If Robbie sees Harriet make 2 paperclips, he infers that she prefers paperclips, so he now believes the value of a paperclip is uniformly distributed between 50⌀ And $1,00, with an average of 75⌀ … In that case, his best plan is to make 90 paperclips with an expected value of $67,50 for Harriet.

(b) If Robbie sees Harriet make 1 of each, he infers that she values paperclips and staples at 50⌀, so the best choice is to make 50 of each.

(c) If Robbie sees Harriet make 2 staples, then by the same argument as in 2(a), he should make 90 staples.

(3) Given this strategy for Robbie, Harriet's best strategy is now somewhat different from the greedy strategy in step 1: if Robbie is going to respond to her making 1 of each by making 50 of each, then she is better off

making 1 of each not just if she is *exactly* indifferent but if she is *anywhere close* to indifferent. In fact, the optimal policy is now to make 1 of each if she values paperclips anywhere between about 44,6∅ and 55,4∅.

(4) Given this new strategy for Harriet, Robbie's strategy remains unchanged. For example, if she chooses 1 of each, he infers that the value of a paperclip is uniformly distributed between 44,6∅ and 55,4∅, with an average of 50∅, so the best choice is to make 50 of each. Because Robbie's strategy is the same as in step 2, Harriet's best response will be the same as in step 3, and we have found the equilibrium.

(13) For a more complete analysis of the off-switch game, see Dylan Hadfield-Menell et al., "The off-switch game," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, ed. Carles Sierra (IJCAI, 2017).

(14) The proof of the general result is quite simple if you don't mind integral signs. Let $P(u)$ be Robbie's prior probability density over Harriet's utility for the proposed action a. Then the value of going ahead with a is

$$EU(a) = \int_{-\infty}^{\infty} P(u) \cdot u\,du = \int_{-\infty}^{0} P(u) \cdot u\,du + \int_{0}^{\infty} P(u) \cdot u\,du$$

(We will see shortly why the integral is split up in this way.) On the other hand, the value of action d, deferring to Harriet, is composed of two parts: if $u > 0$, then Harriet lets Robbie go ahead, so the value is u, but if $u < 0$, then Harriet switches Robbie off, so the value is 0:

$$EU(d) = \int_{-\infty}^{0} P(u) \cdot 0\,du + \int_{0}^{\infty} P(u) \cdot u\,du$$

Comparing the expressions for $EU(a)$ and $EU(d)$, we see immediately that $EU(d) \geq EU(a)$ because the expression for $EU(d)$ has the negative-utility region zeroed out. The two choices have equal value only when the negative region has zero probability — that is, when Robbie is already certain that Harriet likes the proposed action. The theorem is a direct analog

of the well-known theorem concerning the non-negative expected value of information.

(15) Perhaps the next elaboration in line, for the one human–one robot case, is to consider a Harriet who does not yet know her own preferences regarding some aspect of the world, or whose preferences have not yet been formed.

(16) To see how exactly Robbie converges to an incorrect belief, consider a model in which Harriet is slightly irrational, making errors with a probability that diminishes exponentially as the size of error increases. Robbie offers Harriet 4 paperclips in return for 1 staple; she refuses. According to Robbie's beliefs, this is irrational: even at 25∅ Per paperclip and 75∅ per staple, she should accept 4 for 1. Therefore, she must have made a mistake — but this mistake is *much* more likely if her true value is 25∅ than if it is, say, 30∅, because the error costs her a lot more if her value for paperclips is 30∅ … Now Robbie's probability distribution has 25∅ as the most likely value because it represents the smallest error on Harriet's part, with exponentially lower probabilities for values higher than 25∅ … If he keeps trying the same experiment, the probability distribution becomes more and more concentrated close to 25∅ … In the limit, Robbie becomes certain that Harriet's value for paperclips is 25∅.

(17) Robbie could, for example, have a normal (Gaussian) distribution for his prior belief about the exchange rate, which stretches from $-\infty$ to $+\infty$.

(18) For an example of the kind of mathematical analysis that may be needed, see Avrim Blum, Lisa Hellerstein, and Nick Littlestone, "Learning in the presence of finitely or infinitely many irrelevant attributes," *Journal of Computer and System Sciences* 50 (1995): 32–40. Also Lori Dalton, "Optimal Bayesian feature selection," in *Proceedings of the 2013 IEEE Global*

Conference on Signal and Information Processing, ed. Charles Bouman, Robert Nowak, and Anna Scaglione (IEEE, 2013).

(19) Here I am rephrasing slightly a question by Moshe Vardi at the Asilomar Conference on Beneficial AI, 2017.

(20) Michael Wellman and Jon Doyle, "Preferential semantics for goals," in *Proceedings of the 9th National Conference on Artificial Intelligence* (AAAI Press, 1991). This paper draws on a much earlier proposal by Georg von Wright, "The logic of preference reconsidered," *Theory and Decision* 3 (1972): 140–67.

(21) My late Berkeley colleague has the distinction of becoming an adjective. See Paul Grice, *Studies in the Way of Words* (Harvard University Press, 1989).

(22) The original paper on direct stimulation of pleasure centers in the brain: James Olds and Peter Milner, "Positive reinforcement produced by electrical stimulation of septal area and other regions of rat brain," *Journal of Comparative and Physiological Psychology* 47 (1954): 419–27.

(23) Letting rats push the button: James Olds, "Self-stimulation of the brain; its use to study local effects of hunger, sex, and drugs," *Science* 127 (1958): 315–24.

(24) Letting humans push the button: Robert Heath, "Electrical self-stimulation of the brain in man," *American Journal of Psychiatry* 120 (1963): 571–77.

(25) A first mathematical treatment of wireheading, showing how it occurs in reinforcement learning agents: Mark Ring and Laurent Orseau, "Delusion, survival, and intelligent agents," in *Artificial General Intelligence: 4th International Conference*, ed. Jurgen Schmidhuber, Kristinn Thorisson, and Moshe Looks (Springer, 2011). One possible solution to the wireheading problem: Tom Everitt and Marcus Hutter, "Avoiding wireheading with value reinforcement learning," arXiv:1605.03143 (2016).

(26) How it might be possible for an intelligence explosion to occur safely: Benja Fallenstein and Nate Soares, "Vingean reflection: Reliable reasoning for self-improving agents," technical report 2015-2, Machine Intelligence Research Institute, 2015.

(27) The difficulty agents face in reasoning about themselves and their successors: Benja Fallenstein and Nate Soares, "Problems of self-reference in self-improving space-time embedded intelligence," in *Artificial General Intelligence: 7th International Conference*, ed. Ben Goertzel, Laurent Orseau, and Javier Snaider (Springer, 2014).

(28) Showing why an agent might pursue an objective different from its true objective if its computational abilities are limited: Jonathan Sorg, Satinder Singh, and Richard Lewis, "Internal rewards mitigate agent boundedness," in *Proceedings of the 27th International Conference on Machine Learning*, ed. Johannes Furnkranz and Thorsten Joachims (2010), icml.cc/Conferences/2010/papers/icml2010proceedings.zip.

الفصل التاسع: التعقيدات: البشر

(1) Some have argued that biology and neuroscience are also directly relevant. See, for example, Gopal Sarma, Adam Safron, and Nick Hay, "Integrative biological simulation, neuropsychology, and AI safety," arxiv.org/abs/1811.03493 (2018).

(2) On the possibility of making computers liable for damages: Paulius Čerka, Jurgita Grigienė, and Gintarė Sirbikytė, "Liability for damages caused by artificial intelligence," *Computer Law and Security Review* 31 (2015): 376–89.

(3) For an excellent machine-oriented introduction to standard ethical theories and their implications for designing AI systems, see Wendell Wallach and Colin Allen, *Moral Machines: Teaching Robots Right from Wrong* (Oxford University Press, 2008).

(4) The sourcebook for utilitarian thought: Jeremy Bentham, *An Introduction to the Principles of Morals and Legislation* (T. Payne & Son, 1789).

(5) Mill's elaboration of his tutor Bentham's ideas was extraordinarily influential on liberal thought: John Stuart Mill, *Utilitarianism* (Parker, Son & Bourn, 1863).

(6) The paper introducing preference utilitarianism and preference autonomy: John Harsanyi, "Morality and the theory of rational behavior," *Social Research* 44 (1977): 623–56.

(7) An argument for social aggregation via weighted sums of utilities when deciding on behalf of multiple individuals: John Harsanyi, "Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility," *Journal of Political Economy* 63 (1955): 309–21.

(8) A generalization of Harsanyi's social aggregation theorem to the case of unequal prior beliefs: Andrew Critch, Nishant Desai, and Stuart Russell, "Negotiable reinforcement learning for Pareto optimal sequential decision-making," in *Advances in Neural Information Processing Systems 31*, ed. Samy Bengio et al. (2018).

(9) The sourcebook for ideal utilitarianism: G. E. Moore, *Ethics* (Williams & Norgate, 1912).

(10) News article citing Stuart Armstrong's colorful example of misguided utility maximization: Chris Matyszczyk, "Professor warns robots could keep us in coffins on heroin drips," CNET, June 29, 2015.

(11) Popper's theory of negative utilitarianism (so named later by Smart): Karl Popper, *The Open Society and Its Enemies* (Routledge, 1945).

(12) A refutation of negative utilitarianism: R. Ninian Smart, "Negative utilitarianism," *Mind* 67 (1958): 542–43.

(13) For a typical argument for risks arising from "end human suffering" commands, see "Why do we think AI will destroy us?," Reddit,

reddit.com/r/Futurology/comments/38fp6o/why_do_we_think_ai_will_destroy_us.

(14) A good source for self-deluding incentives in AI: Ring and Orseau, “Delusion, survival, and intelligent agents.”

(15) On the impossibility of interpersonal comparisons of utility: W. Stanley Jevons, *The Theory of Political Economy* (Macmillan, 1871).

(16) The utility monster makes its appearance in Robert Nozick, *Anarchy, State, and Utopia* (Basic Books, 1974).

(17) For example, we can fix immediate death to have a utility of 0 and a maximally happy life to have a utility of 1. See John Isbell, “Absolute games,” in *Contributions to the Theory of Games*, vol. 4, ed. Albert Tucker and R. Duncan Luce (Princeton University Press, 1959).

(18) The oversimplified nature of Thanos’s population-halving policy is discussed by Tim Harford, “Thanos shows us how not to be an economist,” *Financial Times*, April 20, 2019. Even before the film debuted, defenders of Thanos began to congregate on the subreddit r/thanosdidnothingwrong/. In keeping with the subreddit’s motto, 350,000 of the 700,000 members were later purged.

(19) On utilities for populations of different sizes: Henry Sidgwick, *The Methods of Ethics* (Macmillan, 1874).

(20) The Repugnant Conclusion and other knotty problems of utilitarian thinking: Derek Parfit, *Reasons and Persons* (Oxford University Press, 1984).

(21) For a concise summary of axiomatic approaches to population ethics, see Peter Eckersley, “Impossibility and uncertainty theorems in AI value alignment,” in *Proceedings of the AAAI Workshop on Artificial Intelligence Safety*, ed. Huáscar Espinoza et al. (2019).

(22) Calculating the long-term carrying capacity of the Earth: Daniel O'Neill et al., "A good life for all within planetary boundaries," *Nature Sustainability* 1 (2018): 88–95.

(23) For an application of moral uncertainty to population ethics, see Hilary Greaves and Toby Ord, "Moral uncertainty about population axiology," *Journal of Ethics and Social Philosophy* 12 (2017): 135–67. A more comprehensive analysis is provided by Will MacAskill, Krister Bykvist, and Toby Ord, *Moral Uncertainty* (Oxford University Press, forthcoming).

(24) Quotation showing that Smith was not so obsessed with selfishness as is commonly imagined: Adam Smith, *The Theory of Moral Sentiments* (Andrew Millar; Alexander Kincaid and J. Bell, 1759).

(25) For an introduction to the economics of altruism, see Serge-Christophe Kolm and Jean Ythier, eds., *Handbook of the Economics of Giving, Altruism and Reciprocity*, 2 vols. (North-Holland, 2006).

(26) On charity as selfish: James Andreoni, "Impure altruism and donations to public goods: A theory of warm-glow giving," *Economic Journal* 100 (1990): 464–77.

(27) For those who like equations: let Alice's intrinsic well-being be measured by w_A and Bob's by w_B. Then the utilities for Alice and Bob are defined as follows:

$$U_A = w_A + C_{AB} w_B$$

$$U_B = w_B + C_{BA} w_A.$$

Some authors suggest that Alice cares about Bob's overall utility U_B rather than just his intrinsic well-being w_B, but this leads to a kind of circularity in that Alice's utility depends on Bob's utility which depends on Alice's utility; sometimes stable solutions can be found but the underlying model can

be questioned. See, for example, Hajime Hori, "Nonpaternalistic altruism and functional interdependence of social preferences," *Social Choice and Welfare* 32 (2009): 59–77.

(28) Models in which each individual's utility is a linear combination of everyone's wellbeing are just one possibility. Much more general models are possible — for example, models in which some individuals prefer to avoid severe inequalities in the distribution of well-being, even at the expense of reducing the total, while other individuals would really prefer that no one have preferences about inequality at all. Thus, the overall approach I am proposing accommodates multiple moral theories held by individuals; at the same time, it doesn't insist that any one of those moral theories is correct or should have much sway over outcomes for those who hold a different theory. I am indebted to Toby Ord for pointing out this feature of the approach.

(29) Arguments of this type have been made against policies designed to ensure equality of outcome, notably by the American legal philosopher Ronald Dworkin. See, for example, Ronald Dworkin, "What is equality? Part 1: Equality of welfare," *Philosophy and Public Affairs* 10 (1981): 185–246. I am indebted to Iason Gabriel for this reference.

(30) Malice in the form of revenge-based punishment for transgressions is certainly a common tendency. Although it plays a social role in keeping members of a community in line, it can be replaced by an equally effective policy driven by deterrence and prevention — that is, weighing the intrinsic harm done when punishing the transgressor against the benefits to the larger society.

(31) Let E_{AB} and P_{AB} be Alice's coefficients of envy and pride respectively, and assume that they apply to the difference in well-being.

Then a (somewhat oversimplified) formula for Alice's utility could be the following:

$$U = w_A + C_{AB} w_B - E_{AB}(w_B - w_A) + P_{AB}(w_A - w_B)$$

$$= (1 + E_{AB} + P_{AB}) w_A + (C_{AB} - E_{AB} - P_{AB}) w_B.$$

Thus, if Alice has positive pride and envy coefficients, they act on Bob's welfare exactly like sadism and malice coefficients: Alice is happier if Bob's welfare is lowered, all other things being equal. In reality, pride and envy typically apply not to differences in well-being but to differences in visible aspects thereof, such as status and possessions. Bob's hard toil in acquiring his possessions (which lowers his overall well-being) may not be visible to Alice. This can lead to the self-defeating behaviors that go under the heading of "keeping up with the Joneses."

(32) On the sociology of conspicuous consumption: Thorstein Veblen, *The Theory of the Leisure Class: An Economic Study of Institutions* (Macmillan, 1899).

(33) Fred Hirsch, *The Social Limits to Growth* (Routledge & Kegan Paul, 1977).

(34) I am indebted to Ziyad Marar for pointing me to social identity theory and its importance in understanding human motivation and behavior. See, for example, Dominic Abrams and Michael Hogg, eds., *Social Identity Theory: Constructive and Critical Advances* (Springer, 1990). For a much briefer summary of the main ideas, see Ziyad Marar, "Social identity," in *This Idea Is Brilliant: Lost, Overlooked, and Underappreciated Scientific Concepts Everyone Should Know*, ed. John Brockman (Harper Perennial, 2018).

(35) Here, I am not suggesting that we necessarily need a detailed understanding of the neural implementation of cognition; what is needed

is a model at the "software" level of how preferences, both explicit and implicit, generate behavior. Such a model would need to incorporate what is known about the reward system.

(36) Ralph Adolphs and David Anderson, *The Neuroscience of Emotion: A New Synthesis* (Princeton University Press, 2018).

(37) See, for example, Rosalind Picard, *Affective Computing*, 2nd ed. (MIT Press, 1998).

(38) Waxing lyrical on the delights of the durian: Alfred Russel Wallace, *The Malay Archipelago: The Land of the Orang-Utan, and the Bird of Paradise* (Macmillan, 1869).

(39) A less rosy view of the durian: Alan Davidson, *The Oxford Companion to Food* (Oxford University Press, 1999). Buildings have been evacuated and planes turned around in mid-flight because of the durian's overpowering odor.

(40) I discovered after writing this chapter that the durian was used for exactly the same philosophical purpose by Laurie Paul, *Transformative Experience* (Oxford University Press, 2014). Paul suggests that uncertainty about one's own preferences presents fatal problems for decision theory, a view contradicted by Richard Pettigrew, Transformative experience and decision theory, *Philosophy and Phenomenological Research* 91 (2015): 766–74. Neither author refers to the early work of Harsanyi, Games with incomplete information, Parts I–III, or Cyert and de Groot, Adaptive utility.

(41) An initial paper on helping humans who don't know their own preferences and are learning about them: Lawrence Chan et al., "The assistive multi-armed bandit," in *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, ed. David Sirkin et al. (IEEE, 2019).

(42) Eliezer Yudkowsky, in *Coherent Extrapolated Volition* (Singularity Institute, 2004), lumps all these aspects, as well as plain inconsistency, under the heading of *muddle*—a term that has not, unfortunately, caught on.

(43) On the two selves who evaluate experiences: Daniel Kahneman, *Thinking, Fast and Slow* (Farrar, Straus & Giroux, 2011).

(44) Edgeworth's hedonimeter, an imaginary device for measuring happiness moment to moment: Francis Edgeworth, *Mathematical Psychics: An Essay on the Application of Mathematics to the Moral Sciences* (Kegan Paul, 1881).

(45) A standard text on sequential decisions under uncertainty: Martin Puterman, *Markov Decision Processes*: *Discrete Stochastic Dynamic Programming* (Wiley, 1994).

(46) On axiomatic assumptions that justify additive representations of utility over time: Tjalling Koopmans, "Representation of preference orderings over time," in *Decision and Organization*, ed. C. Bartlett McGuire, Roy Radner, and Kenneth Arrow (North-Holland, 1972).

(47) The 2019 humans (who might, in 2099, be long dead or might just be the earlier selves of 2099 humans) might wish to build the machines in a way that respects the 2019 preferences of the 2019 humans rather than pandering to the undoubtedly shallow and ill-considered preferences of humans in 2099. This would be like drawing up a constitution that disallows any amendments. If the 2099 humans, after suitable deliberation, decide they wish to override the preferences built in by the 2019 humans, it seems reasonable that they should be able to do so. After all, it is they and their descendants who have to live with the consequences.

(48) I am indebted to Wendell Wallach for this observation.

(49) An early paper dealing with changes in preferences over time: John Harsanyi, "Welfare economics of variable tastes," *Review of Economic Studies* 21 (1953): 204–13. A more recent (and somewhat technical) survey

is provided by Franz Dietrich and Christian List, "Where do preferences come from?," *International Journal of Game Theory* 42 (2013): 613–37. See also Laurie Paul, *Transformative Experience* (Oxford University Press, 2014), and Richard Pettigrew, "Choosing for Changing Selves," philpapers .org/archive/PETCFC.pdf.

(50) For a rational analysis of irrationality, see Jon Elster, *Ulysses and the Sirens: Studies in Rationality and Irrationality* (Cambridge University Press, 1979).

(51) For promising ideas on cognitive prostheses for humans, see Falk Lieder, "Beyond bounded rationality: Reverse-engineering and enhancing human intelligence" (PhD thesis, University of California, Berkeley, 2018).

الفصل العاشر: هل حُلَّت المشكلة؟

(1) On the application of assistance games to driving: Dorsa Sadigh et al., "Planning for cars that coordinate with people," *Autonomous Robots* 42 (2018): 1405–26.

(2) Apple is, curiously, absent from this list. It does have an AI research group and is ramping up rapidly. Its traditional culture of secrecy means that its impact in the marketplace of ideas is quite limited so far.

(3) Max Tegmark, interview, *Do You Trust This Computer?*, directed by Chris Paine, written by Mark Monroe (2018).

(4) On estimating the impact of cybercrime: "Cybercrime cost $600 billion and targets banks first," *Security Magazine*, February 21, 2018.

الملحق «أ»: البحث عن حلول

(1) The basic plan for chess programs of the next sixty years: Claude Shannon, "Programming a computer for playing chess," *Philosophical*

Magazine, 7th ser., 41 (1950): 256–75. Shannon's proposal drew on a centuries-long tradition of evaluating chess positions by adding up piece values; see, for example, Pietro Carrera, *Il gioco degli scacchi* (Giovanni de Rossi, 1617).

(2) A report describing Samuel's heroic research on an early reinforcement learning algorithm for checkers: Arthur Samuel, "Some studies in machine learning using the game of checkers," *IBM Journal of Research and Development* 3 (1959): 210–29.

(3) The concept of rational metareasoning and its application to search and game playing emerged from the thesis research of my student Eric Wefald, who died tragically in a car accident before he could write up his work; the following appeared posthumously: Stuart Russell and Eric Wefald, *Do the Right Thing: Studies in Limited Rationality* (MIT Press, 1991). See also Eric Horvitz, "Rational metareasoning and compilation for optimizing decisions under bounded resources," in *Computational Intelligence, II: Proceedings of the International Symposium,* ed. Francesco Gardin and Giancarlo Mauri (North-Holland, 1990); and Stuart Russell and Eric Wefald, "On optimal game-tree search using rational meta-reasoning," in *Proceedings of the 11th International Joint Conference on Artificial Intelligence*, ed. Natesa Sridharan (Morgan Kaufmann, 1989).

(4) Perhaps the first paper showing how hierarchical organization reduces the combinatorial complexity of planning: Herbert Simon, "The architecture of complexity," *Proceedings of the American Philosophical Society* 106 (1962): 467–82.

(5) The canonical reference for hierarchical planning is Earl Sacerdoti, "Planning in a hierarchy of abstraction spaces," *Artificial Intelligence* 5 (1974): 115–35. See also Austin Tate, "Generating project networks," in *Proceedings of the 5th International Joint Conference on Artificial Intelligence*, ed. Raj Reddy (Morgan Kaufmann, 1977).

(6) A formal definition of what high-level actions do: Bhaskara Marthi, Stuart Russell, and Jason Wolfe, “Angelic semantics for high-level actions,” in *Proceedings of the 17th International Conference on Automated Planning and Scheduling*, ed. Mark Boddy, Maria Fox, and Sylvie Thiebaux (AAAI Press, 2007).

الملحق «ب»: المعرفة والمنطق

(1) This example is unlikely to be from Aristotle, but may have originated with Sextus Empiricus, who lived probably in the second or third century CE.

(2) The first algorithm for theorem-proving in first-order logic worked by reducing firstorder sentences to (very large numbers of) propositional sentences: Martin Davis and Hilary Putnam, “A computing procedure for quantification theory,” *Journal of the ACM* 7 (1960): 201–15.

(3) An improved algorithm for propositional inference: Martin Davis, George Logemann, and Donald Loveland, “A machine program for theorem-proving,” *Communications of the ACM* 5 (1962): 394–97.

(4) The satisfiability problem — deciding whether a collection of sentences is true in *some* world — is NP-complete. The reasoning problem — deciding whether a sentence follows from the known sentences — is co-NP-complete, a class that is thought to be harder than NP-complete problems.

(5) There are two exceptions to this rule: no repetition (a stone may not be played that returns the board to a situation that existed previously) and no suicide (a stone may not be placed such that it would immediately be captured — for example, if it is already surrounded).

(6) The work that introduced first-order logic as we understand it today (*Begriffsschrift* means “concept writing”): Gottlob Frege, *Begriffsschrift,*

eine der arithmetischen nachgebildete Formelsprache des reinen Denkens (Halle, 1879). Frege's notation for first-order logic was so bizarre and unwieldy that it was soon replaced by the notation introduced by Giuseppe Peano, which remains in common use today.

(7) A summary of Japan's bid for supremacy through knowledge-based systems: Edward Feigenbaum and Pamela McCorduck, *The Fifth Generation: Artificial Intelligence and Japan's Computer Challenge to the World* (Addison-Wesley, 1983).

(8) The US efforts included the Strategic Computing Initiative and the formation of the Microelectronics and Computer Technology Corporation (MCC). See Alex Roland and Philip Shiman, *Strategic Computing: DARPA and the Quest for Machine Intelligence, 1983–1993* (MIT Press, 2002).

(9) A history of Britain's response to the re-emergence of AI in the 1980s: Brian Oakley and Kenneth Owen, *Alvey: Britain's Strategic Computing Initiative* (MIT Press, 1990).

(10) The origin of the term *GOFAI*: John Haugeland, *Artificial Intelligence: The Very Idea* (MIT Press, 1985).

(11) Interview with Demis Hassabis on the future of AI and deep learning: Nick Heath, "Google DeepMind founder Demis Hassabis: Three truths about AI," *TechRepublic*, September 24, 2018.

الملحق «ج»: عدم اليقين والاحتمال

(1) Pearl's work was recognized by the Turing Award in 2011.

(2) Bayes nets in more detail: Every node in the network is annotated with the probability of each possible value, given each possible combination of values for the node's *parents* (that is, those nodes that point to it). For example, the probability that $Doubles_{12}$ has value *true* is 1,0 when D_1 and D_2 have the same value, and 0,0 otherwise. A possible world

is an assignment of values to all the nodes. The probability of such a world is the product of the appropriate probabilities from each of the nodes.

(3) A compendium of applications of Bayes nets: Olivier Pourret, Patrick Naïm, and Bruce Marcot, eds., *Bayesian Networks: A Practical Guide to Applications* (Wiley, 2008).

(4) The basic paper on probabilistic programming: Daphne Koller, David McAllester, and Avi Pfeffer, “Effective Bayesian inference for stochastic programs,” in *Proceedings of the 14th National Conference on Artificial Intelligence* (AAAI Press, 1997). For many additional references, see probabilistic-programming.org.

(5) Using probabilistic programs to model human concept learning: Brenden Lake, Ruslan Salakhutdinov, and Joshua Tenenbaum, “Human-level concept learning through probabilistic program induction,” *Science* 350 (2015): 1332–38.

(6) For a detailed description of the seismic monitoring application and associated probability model, see Nimar Arora, Stuart Russell, and Erik Sudderth, “NET-VISA: Network processing vertically integrated seismic analysis,” *Bulletin of the Seismological Society of America* 103 (2013): 709–29.

(7) News article describing one of the first serious self-driving car crashes: Ryan Randazzo, “Who was at fault in self-driving Uber crash? Accounts in Tempe police report disagree,” *Republic* (azcentral.com), March 29, 2017.

الملحق «د»: التعلم من التجربة

(1) The foundational discussion of inductive learning: David Hume, *Philosophical Essays Concerning Human Understanding* (A. Millar, 1748).

(2) Leslie Valiant, “A theory of the learnable,” *Communications of the ACM* 27 (1984): 1134–42. See also Vladimir Vapnik, *Statistical Learning*

Theory (Wiley, 1998). Valiant's approach concentrated on computational complexity, Vapnik's on statistical analysis of the learning capacity of various classes of hypotheses, but both shared a common theoretical core connecting data and predictive accuracy.

(3) For example, to learn the difference between the "situational superko" and "natural situational superko" rules, the learning algorithm would have to try repeating a board position that it had created previously by a pass rather than by playing a stone. The results would be different in different countries.

(4) For a description of the ImageNet competition, see Olga Russakovsky et al., "ImageNet large scale visual recognition challenge," *International Journal of Computer Vision* 115 (2015): 211–52.

(5) The first demonstration of deep networks for vision: Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, ed. Fernando Pereira et al. (2012).

(6) The difficulty of distinguishing over one hundred breeds of dogs: Andrej Karpathy, "What I learned from competing against a ConvNet on ImageNet," *Andrej Karpathy Blog*, September 2, 2014.

(7) Blog post on inceptionism research at Google: Alexander Mordvintsev, Christopher Olah, and Mike Tyka, "Inceptionism: Going deeper into neural networks," *Google AI Blog*, June 17, 2015. The idea seems to have originated with J. P. Lewis, "Creation by refinement: A creativity paradigm for gradient descent learning networks," in *Proceedings of the IEEE International Conference on Neural Networks* (IEEE, 1988).

(8) News article on Geoff Hinton having second thoughts about deep networks: Steve LeVine, "Artificial intelligence pioneer says we need to start over," *Axios*, September 15, 2017.

(9) A catalog of shortcomings of deep learning: Gary Marcus, “Deep learning: A critical appraisal,” arXiv:1801.00631 (2018).

(10) A popular textbook on deep learning, with a frank assessment of its weaknesses: François Chollet, *Deep Learning with Python* (Manning Publications, 2017).

(11) An explanation of explanation-based learning: Thomas Dietterich, “Learning at the knowledge level,” *Machine Learning* 1 (1986): 287–315.

(12) A superficially quite different explanation of explanation-based learning: John Laird, Paul Rosenbloom, and Allen Newell, “Chunking in Soar: The anatomy of a general learning mechanism,” *Machine Learning* 1 (1986): 11–46.


هنداوي